

# **Visual Monocular SLAM for Minimally Invasive Surgery and its Application to Augmented Reality**

Thesis presented by

**Nader Mahmoud Elshahat Elsayed ALI**

**Defende publicly on: 19 June 2018**

Submitted to

**University of Strasbourg**

**Universidad de Zaragoza**

For obtaining the degree of

**PhD** (Doctor of Philosophy)

## **Thesis Directors:**

**Christophe Doignon**

Professor, University of Strasbourg, France

**Jose Maria Martinez Montiel**

Professor, University of Zaragoza, Spain

---

## **Reviewers:**

**Guillaume Morel**

Professor, Pierre-and-Marie-Curie University, Paris, France

**Danail Stoyanov**

Associate Professor, University College London, England

---

## **Examiner:**

**Marie-Odile Berger**

Research Fellow,  
Head of MAGRIT INRIA Nancy Grand Est (Loria), France

---

## **Invited Member:**

**Luc Soler**

Professor, Scientific director at IRCAD, Strasbourg, France



# **Localisation et Cartographie Simultanées par Vision Monoculaire pour la Réalité Médicale Augmentée**

Présentée par :

**Nader Mahmoud Elshahat Elsayed ALI**

**Soutenue le: 19 Juin 2018**

Pour obtenir le grade de

**Docteur de l'université de Strasbourg**  
**Docteur de l'université de Zaragoza**

## **Directeurs de Thèse:**

**Christophe Doignon**

Professeur, Université de Strasbourg, France

**Jose Maria Martinez Montiel**

Professeur, Université de Saragosse, Espagne

---

## **Rapporteurs:**

**Guillaume Morel**

Professeur, Université Pierre et Marie Curie, Paris, France

**Danail Stoyanov**

Professeur associé, Collège universitaire de Londres, Angleterre

---

## **Examineur:**

**Marie-Odile Berger**

Chargée de recherche, responsable équipe MAGRIT,  
INRIA Nancy Grand Est (Loria), France

---

## **Membre Invité:**

**Luc Soler**

Professeur, Directeur scientifique de l'IRCAD Strasbourg, France

## Copyright

You are free to **Share** — copy and redistribute the material in any medium or format. Under the following terms: **Attribution** — you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. **NonCommercial** — you may not use the material for commercial purposes. **NoDerivatives** — if you remix, transform, or build upon the material, you may not distribute the modified material.





To my beloved parents, sisters, wife and sons



# Acknowledgements

First and foremost, I would like to thank the almighty Allah, for his grace and blessings during my doctoral studies. It has been a tough 3-years PhD experience, however it would not be possible without the support of many individuals and organizations. The credit goes to people mentioned below, without any order.

This thesis would not be possible without the support of **IRCAD France** and **IHU**. I would like to thank everyone there for facilitating the procedures to do the required experiments. I wish to express my sincere gratitude to Prof. **Luc Soler**, scientific director of IRCAD, for the funding and for the opportunity to do my PhD studies at such nice place. I'm also grateful to Dr. **Stéphane Nicolau** for being there for me in an early stage of this research. Thanks for your fruitful discussions and pertinent guidance in the first early stages. Thanks also for Dr. **Óscar García Grasa** who has been important contributor in the first stage of this thesis. This thesis would not also be possible without the support of Prof. **Christophe Doignon**, my thesis director, who has always been nothing but kind, supportive and positive about my work.

I also would like to thank Dr. **Alexandre Hostettler** for his efforts in managing funding issues to attend conferences, summer schools and most importantly to schedule experiment sessions inside the operating room, a task which is rather difficult. I also would like to express my gratitude to **Toby Collins** for his participation at late stage of this research. He has enlighten my writing skills, if I had any, and the scientific discussion with him was constructive.

I'm also grateful to **UNIZAR** for the nice hospitality at their labs, and I would like to thank every one there for offering the required means (funded by the Spanish government DPI2015-67275-P and Aragonese DGAT04-FSE) to continue the research work during my research stay. I would like to thank Prof. **José María Martínez Montiel**, my thesis director, who helped me to develop insights into visual SLAM problems. I'm so grateful for his support, guidance, invaluable discussions, and his painstaking efforts in proofreading paper drafts. I'm also grateful to Dr. **Alejo Concha** for motivating me and for his assistance to get in touch with dense SLAM. I also thank all my friends at UNIZAR labs for the warm welcome and making me feel as one of them during my stay. I wish all of you all the best and success in your careers.

Last but not least, I would like to thank my family: parents and sisters for always being there for me and keep me grounded. They are the constant and solid support that I always count on. A strong motivation during this journey was to see them proud. A special thanks goes to my wife and my son during this 3-years roll coaster ride. Thanks for struggling with me and making this journey so enjoyable.



# Abstract

Recovering 3D information of intra-operative endoscopic images together with the relative endoscope camera position are fundamental blocks towards accurate guidance and navigation in image-guided surgery. They allow augmented reality overlay of pre-operative models, which are readily available from different imaging modalities. This thesis provides a systematic approach for estimating these two pieces of information based on a pure vision Simultaneous Localization And Mapping (SLAM). SLAM goal is localizing a camera sensor, in real-time, within a map (3D reconstruction) of the environment that is also built online. It enables markerless camera tracking, where it uses only information from RGB images of a standard monocular camera.

The preliminary work in this thesis has presented a sparse SLAM solution for real time and accurate intra-operative visualization of patient's pre-operative models over the patient skin. We proposed a non-invasive registration and visualization pipeline that requires minimal interactions from medical staff and runs solely on a commodity Tablet-PC with a build-in camera. Subsequently, we directed our focus to endoscopy, which is very challenging for monocular 3D reconstruction and endoscope camera tracking. We have addressed the utilization of the state-of-the-art sparse SLAM, and achieved a remarkable tracking performance. Thus, it was our second contribution to propose a pairwise dense reconstruction algorithm that exploits the initial SLAM exploration phase and accurately provides a pairwise dense reconstruction of the surgical scene.

A further contribution is an extension of state-of-the-art sparse SLAM with a novel dense multi-view stereo-like approach to perform live dense reconstructions and hence eliminates the wait for the abdominal cavity exploration. We decouple the dense reconstruction from the camera trajectory estimation, resulting in a system that combines the accuracy and robustness of feature-based SLAM with the more complete reconstruction of direct SLAM methods. The proposed system can cope with challenging lighting conditions and poor/repetitive textures in endoscopy at an affordable time budget using modern GPU. The proposed system has been validated and evaluated on real porcine sequences of abdominal cavity exploration and showed a superior performance to other dense SLAM methods in terms of accuracy, density, and computation times. It has been also tested on different in-door sequences and showed a promising reconstructions results.

The proposed solutions in this thesis have been validated on real porcine in-vivo and ex-vivo sequences from different datasets and have proved to be fast and do not need any external tracking hardware nor significant intervention from medical staff, other than moving the Tablet-PC or the endoscope. They therefore can be integrated easily into the current surgical workflow.



# Contents

<b>Acknowledgements</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Minimally Invasive Surgery (MIS)	1
1.1.1 Background	1
1.1.2 Medical challenges	2
1.1.3 Augmented reality, next step in MIS	3
1.2 Our goal: dense monocular visual SLAM for MIS	3
1.2.1 Why vision-based endoscope tracking?	4
1.2.2 Why dense?	4
1.2.3 Why dense SLAM in MIS is challenging?	5
1.3 Contributions	5
1.4 Publications	6
<b>2 Literature Review</b>	<b>9</b>
2.1 Introduction	9
2.2 Monocular visual cue	10
2.2.1 Sparse approaches	10
2.2.1.1 Factorization based	10
2.2.1.2 Bundle adjustment based	11
2.2.1.3 SLAM	12
2.2.2 Dense approaches	14
2.2.2.1 Shading cue	15
2.2.2.2 Multi-view stereo	15
2.2.2.3 Feature-based tracking and dense mapping	16
2.2.2.4 Dense tracking and mapping	17
2.3 Stereo visual cue	18
2.3.1 Only dense	18
2.3.2 Stereo SLAM	19
2.4 Dataset	20
2.5 Discussion	21

<b>3</b>	<b>On-Patient See-through Augmented Reality based on visual SLAM</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Related work . . . . .	25
3.3	Method overview . . . . .	25
3.4	Method description . . . . .	27
3.4.1	SLAM architecture . . . . .	27
3.4.1.1	Preliminaries . . . . .	27
3.4.1.2	Camera tracking . . . . .	29
3.4.1.3	Mapping . . . . .	30
3.4.1.4	Bootstrapping . . . . .	30
3.4.1.5	Camera relocation . . . . .	31
3.4.2	Alignment of pre-operative model to SLAM map . . . . .	31
3.4.3	See-through AR . . . . .	33
3.5	Experimental results . . . . .	34
3.5.1	Volunteers experiments: computation time evaluation . . . . .	34
3.5.2	Accuracy evaluation . . . . .	36
3.5.2.1	Data acquisition . . . . .	36
3.5.2.2	Registration accuracy on pigs . . . . .	37
3.5.2.3	Registration accuracy on phantom . . . . .	38
3.6	Conclusion . . . . .	39
<b>4</b>	<b>SLAM Based Quasi Dense Reconstruction For MIS</b>	<b>41</b>
4.1	Introduction . . . . .	42
4.2	ORB-SLAM overview . . . . .	43
4.2.1	Camera tracking . . . . .	43
4.2.2	Mapping . . . . .	44
4.2.3	Camera relocation . . . . .	45
4.3	ORB-SLAM in MIS scenes . . . . .	45
4.3.1	Parameters tunings . . . . .	45
4.3.2	ORB-SLAM in action . . . . .	47
4.4	Densified discrete mapping . . . . .	49
4.5	Quasi dense pairwise reconstruction . . . . .	51
4.5.1	Approach overview . . . . .	52
4.5.2	Frame pre-processing . . . . .	52
4.5.3	Building keyframe neighborhood graph . . . . .	53
4.5.4	Feature based densification . . . . .	53
4.5.5	Featureless depth propagation . . . . .	54
4.5.6	Outliers removal and denoising . . . . .	55
4.6	Experimental evaluation . . . . .	57
4.6.1	Data acquisition . . . . .	57
4.6.2	Quantitative analysis . . . . .	58
4.6.2.1	Sparse reconstruction error . . . . .	59
4.6.2.2	Dense reconstruction error . . . . .	61
4.6.3	Tracking robustness . . . . .	62
4.6.4	Tuning details and computation cost . . . . .	63

4.6.5	AR superimposition of intra-operative CT models . . . . .	64
4.6.6	Performance on indoor sequences . . . . .	65
4.7	Conclusion . . . . .	67
<b>5</b>	<b>Live Tracking and Dense Reconstruction For MIS</b>	<b>69</b>
5.1	Introduction . . . . .	70
5.2	Approach overview . . . . .	71
5.3	Frames cluster selection for dense reconstruction and cluster bundle adjustment	72
5.4	Reconstruction of a keyframe’s depth map . . . . .	73
5.4.1	The variational formulation . . . . .	73
5.4.2	ZNCC data term . . . . .	73
5.4.3	The regularizer . . . . .	74
5.4.4	Initialization . . . . .	75
5.4.5	Energy minimization . . . . .	76
5.4.5.1	Solution . . . . .	76
5.5	Live alignment of keyframe depth maps . . . . .	78
5.6	Experimental Results . . . . .	78
5.6.1	Benchmark hardware and compared methods . . . . .	78
5.6.2	Datasets . . . . .	78
5.6.3	Quantitative evaluation using dense stereo . . . . .	79
5.6.3.1	Evaluation metrics . . . . .	80
5.6.3.2	Results analysis . . . . .	81
5.6.3.3	The influence of the regularizer and number of images in the cluster . . . . .	84
5.6.3.4	Robustness of ZNCC versus SAD for the data term . . . . .	86
5.6.3.5	Proposed system versus dense SfM . . . . .	87
5.6.4	Qualitative evaluation on patient data . . . . .	87
5.6.5	Free parameters tuning . . . . .	88
5.6.6	Processing time . . . . .	88
5.6.7	Augmented reality annotations . . . . .	89
5.6.8	Performance on indoor sequences . . . . .	90
5.7	Conclusion . . . . .	92
<b>6</b>	<b>Conclusions and Future Directions</b>	<b>95</b>
6.1	Conclusions . . . . .	95
6.2	Future directions . . . . .	96
6.2.1	Improvements to our dense SLAM system . . . . .	97
6.2.2	Clinical trials . . . . .	97
6.2.3	Is dense MIS SLAM the holy grail for AR? . . . . .	98
6.2.4	Are the rigid assumptions enough? . . . . .	98
6.2.5	Deep learning . . . . .	98
6.2.6	Comprehensive and diverse dataset . . . . .	99
6.2.7	Robotized MIS . . . . .	99



# List of Figures

1.1	Typical MIS. . . . .	2
2.1	Taxonomy of related works. . . . .	10
3.1	Port positioning with AR guidance during trans-thoracic minimally invasive hepatectomy (Hallet et al., 2015). (a) Pre-operative trocar placement planning. (b) and (c) Marking of the chosen port site. . . . .	24
3.2	System workflow. . . . .	26
3.3	SLAM architecture. . . . .	28
3.4	Pre-operative model alignment within SLAM map. . . . .	31
3.5	AR insertion. (a) Tablet-PC camera frame with projected (red) and matched (green) map points. (b) Virtual 3D scene including the registered pre-operative model. (c) Virtual camera image. (d) Fused AR image. . . . .	33
3.6	Experiment on first volunteer. (a) Real Tablet-PC image. (b) Skin registration with anchor points (in blue) and map points. (c-f) Skin AR overlays over frames from different points of view. . . . .	35
3.7	Registration of transparent skin, liver, left kidney and right kidney on the body of the second volunteer from different points of view. . . . .	35
3.8	Experiments on pigs. (a) Nine radio-opaque markers were attached to the surface of the pig. (b) Pre-operative model composed of skin, bones, liver, left kidney and right kidney overlaid on an image of the first pig. (c) Lateral view of skin registration on the second pig. (d) Keyframe locations during camera motion around one pig. . . . .	36
3.9	Evolution of the average distances error between $\mathbf{l}_i$ and $\mathbf{L}_i$ (in mm) over all frames of one video sequence in case of breath-hold. (a) The average distances between the 5 markers used in the registration. (b) The average distances of the remaining 4 markers. . . . .	38
3.10	Experiments on phantom. (a) Markers in blue were used for the registration. (b) and (c) Liver AR overlay with partial scene occlusion. . . . .	38
4.1	Image samples from training database used for DBoW2. . . . .	47

4.2	ORB-SLAM performance. (a-c) Image samples with projected map points in green. (d-e) Camera tracking during exploration, current endoscope pose is shown as a green frustum during inhale (d) and exhale (e), respectively. (f) Reconstructed map and keyframes locations (blue frustum), it corresponds to endoscope tip trajectory. . . . .	48
4.3	Gastroscopy sequence. (a) Esophagus with tracked points. (b) Reconstructed map, keyframes, and current endoscope location. . . . .	48
4.4	Relocalization. (a) Consecutive stages from left to right: successful tracking while observing the liver, tracking loss due to laparoscope extraction, laparoscope re-insertion towards the spleen, and relocating laparoscope pose and resume tracking. (b,c) The arrows refer to the laparoscope locations before and after relocalization. . . . .	49
4.5	Epipolar guided search. $KF_i$ is the current keyframe and $KF_j$ is its neighbored keyframe . . . . .	50
4.6	Improved scene mapping. (a) Image sample with original ORB-SLAM points projected in yellow. (b) Original ORB-SLAM map. (c) Map when activating cross correlation search for correspondences search. (d,e) Incremental reconstruction during laparoscope exploration. (f) Final map with points shown in original RGB intensities. . . . .	51
4.7	Quasi dense pairwise reconstruction pipeline. . . . .	52
4.8	Feature based densification. . . . .	54
4.9	Featureless depth propagation. (a,b) Two keyframes with corresponding features. (c) Disparity map, where darker pixels are closer. (d,e) Depth propagation in SLAM map. . . . .	55
4.10	Surface denoising. . . . .	56
4.11	Denoised reconstructed map from private (a,b) and public (c,d) datasets. . . . .	57
4.12	Data acquisition. (a) Video recording. (b) CT acquisition while laparoscope is fixed and its tip inside the abdominal cavity. (c) Complete CT surface of pig abdominal cavity. . . . .	58
4.13	Distances distributions of tuned ORB-SLAM map. . . . .	59
4.14	Alignment of tuned ORB-SLAM map to CT model. (a,b) Alignment from two points of view with the visible part of CT surface in laparoscope images. (d) Outliers rejection, where yellow, white and red points are CT model points, inliers and outliers map points, respectively. . . . .	60
4.15	Distances distributions of discrete densified map. . . . .	60
4.16	Alignment of discrete densified map to CT model. . . . .	61
4.17	Distances distributions of dense map. . . . .	61
4.18	Alignment of quasi dense map to CT. . . . .	61
4.19	Endoscope tracking and mapping during partial scene occlusions and deformations. . . . .	62
4.20	Markerless AR overlay of liver hepatic vein. (a,b) Alignment of hepatic vein (blue), liver(green) and abdominal wall(yellow). (c-f) Image samples of AR rendering during exploration . . . . .	64
4.21	Pairwise dense reconstruction of different indoor sequences from public dataset (Sturm et al., 2012). . . . .	66
5.1	System Architecture. . . . .	71

5.2	Cost volume construction of reference keyframe $I_r$ with relative pose $\mathbf{T}_{r,w}$ with respect to SLAM map $w$ and $n$ neighbor frames with relative pose $\mathbf{T}_{i_n,r}$ with respect to $I_r$ . Each pixel in $I_r$ has an associated row of entries in cost volume (shown in red) that store the averageZNCC cost computed for the corresponding $\rho \in [\beta_{min}\rho_{min}, \beta_{max}\rho_{max}]$ . . . . .	75
5.3	Sample frames of the different laparoscope porcine sequences used from public Mountney et al., 2010b and private datasets. . . . .	79
5.4	Monocular (green) and stereo (textured) reconstruction after scale alignment. Stereo cameras are show in red, and cluster of frames in grey. . . . .	80
5.5	Euclidean distances between the stereo and the monocular dense maps. . . . .	80
5.6	Error evolution between selected keyframes for dense reconstruction. . . . .	82
5.7	Incremental dense reconstruction of proposed system and LSD-SLAM on different sequences visualized as point clouds. SLAM keyframes and points are colored in blue and in red, respectively. The selected keyframes used for the dense reconstruction and frames cluster are colored in red and grey, respectively. The green frustum shows the current laparoscope pose. . . . .	84
5.8	Effect of the regularizer and the number of processed images in the cluster. . . .	85
5.9	Initial depth maps obtained when using ZNCC (left) and SAD (right) for the data term in the variational problem. . . . .	86
5.10	Dense SfM (Langguth et al., 2016). . . . .	87
5.11	Reconstruction results on patient liver sequence. (a) Image sample, (b) Sparse ORB-SLAM reconstruction. (c-d) Dense reconstruction of liver surface by our system, from different directions. . . . .	88
5.12	Using dense surface for AR annotations. (a,b) Adding AR annotation on the reconstructed dense surface, red frustum is virtual camera placed at estimated laparoscope pose. (c) AR annotation at estimated laparoscope pose in (a,b). (d-f) AR view during laparoscope exploration in right-to-left direction . . . . .	90
5.13	Performance on different indoor scenes from public dataset (Sturm et al., 2012). . . . .	91



# List of Tables

2.1	Publicly available MIS Datasets. . . . .	21
3.1	$\overline{FRE}$ and $\overline{TRE}$ (in mm) of the four pigs sequences recorded during breath-hold. . . . .	37
3.2	$\overline{FRE}$ and $\overline{TRE}$ (in mm) of the four pigs sequences recorded during breathing. . . . .	37
3.3	$\overline{FRE}$ and $\overline{TRE}$ (in mm) after processing the whole phantom sequence. . . . .	39
4.1	Average computational cost of ORB-SLAM + dense discrete mapping average . . . . .	63
4.2	Average computational cost of quasi dense pariwise reconstruction. . . . .	63
4.3	Tuning parameters used for dense pairwise reconstruction. . . . .	63
4.4	Average reconstruction error with respect to RGB-D sensor. . . . .	67
5.1	Average reconstruction error with respect to stereo methods (Hirschmuller, 2008, Chang et al., 2013). . . . .	81
5.2	RMSE errors in mm with respect to Chang et al., 2013, when use ZNCC and SAD for the data term in the variational problem. . . . .	86
5.3	Parameters tuning. . . . .	88
5.4	Average processing time (in seconds). . . . .	89
5.5	Average reconstruction error with respect to RGB-D sensor. . . . .	92



# List of Abbreviations

<b>MIS</b>	Minimally Invasive Surgery
<b>CO<sub>2</sub></b>	Carbon Dioxide Gas
<b>GI</b>	Gastrointestinal
<b>FOV</b>	Field Of View
<b>CT</b>	Computerized Tomography
<b>MRI</b>	Magnetic Resonance Imaging
<b>AR</b>	Augmented Reality
<b>SLAM</b>	Simultaneous Localization And Mapping
<b>SfM</b>	Structure from Motion
<b>SVD</b>	Singular Value Decomposition
<b>BA</b>	Bundle Adjustment
<b>CEA</b>	Competitive Evolutionary Agent
<b>TECAB</b>	Totally Endoscopic Coronary Artery Bypass
<b>ASKC</b>	Adaptive Scale Kernel Consensus
<b>SIFT</b>	Scale-Invariant Feature Transform
<b>ORB</b>	Oriented FAST and Rotated BRIEF
<b>FAST</b>	Features from Accelerated Segment Test
<b>SURF</b>	Speeded up robust features
<b>KF</b>	Kalman Filter
<b>EKF</b>	Extended Kalman Filter
<b>RANSAC</b>	Random Sample Consensus
<b>JCBB</b>	Joint Compatibly Branch and Bound
<b>RLR</b>	Randomized List Relocalization (RLR)
<b>PTAM</b>	Parallel Tracking And Mapping
<b>DTAM</b>	Dense Tracking and Mapping
<b>RALP</b>	Robot-Assisted Laparoscopic Prostatectomy
<b>SfS</b>	Shape from Shading
<b>ICP</b>	Iterative Closest Point
<b>MVS</b>	Multi-View Stereo
<b>ZNCC</b>	Zero-mean Normalized Cross Correlation
<b>HRM</b>	Hybrid Recursive Matching
<b>HMA</b>	Hierarchical Multi-Affine
<b>ESBS</b>	Endoscopic Skull Base Surgery

<b>VTK</b>	Visualization Toolkit
<b>FRE</b>	Fiducial Registration Error
<b>TRE</b>	Target Registration Error
<b>DBoW2</b>	Bag Of Binary Words
<b>PnP</b>	Perspective-n-Point
<b>DLT</b>	Direct Linear Transformation
<b>HSV</b>	Hue, Saturation, Value color space
<b>RMSE</b>	Root Mean Square Error
<b>SAD</b>	Sum of Absolute Difference
<b>SSD</b>	Sum of Squared Difference
<b>TV</b>	Total Variation



# Chapter 1

## Introduction

### Contents

<b>1.1</b>	<b>Minimally Invasive Surgery (MIS)</b>	<b>1</b>
1.1.1	Background	1
1.1.2	Medical challenges	2
1.1.3	Augmented reality, next step in MIS	3
<b>1.2</b>	<b>Our goal: dense monocular visual SLAM for MIS</b>	<b>3</b>
1.2.1	Why vision-based endoscope tracking?	4
1.2.2	Why dense?	4
1.2.3	Why dense SLAM in MIS is challenging?	5
<b>1.3</b>	<b>Contributions</b>	<b>5</b>
<b>1.4</b>	<b>Publications</b>	<b>6</b>

## 1.1 Minimally Invasive Surgery (MIS)

### 1.1.1 Background

Intervention techniques have gained a substantial popularity over the past decade. Surgeons perform such interventions by manipulating an endoscope and surgical instruments (cf. Figure 1.1(a)). The motion of the endoscope and the instruments is executed either by the surgeon, assistant or a surgical robot, e.g: da Vinci system. Laparoscopy is an MIS procedure that is used to examine different organs inside the patient's abdominal cavity using long, thin tube laparoscope (cf. Figure 1.1(b)), during the procedure, the surgeon makes several small incisions in the patient skin and the laparoscope is inserted through one of these incisions to the abdominal cavity. The other surgical instruments are passed through the other openings. Both laparoscope and instruments are inserted through a 5-10mm cannula-shaped input ports called *trocars*. The laparoscope is considered as the surgeon's eye to observe the abdominal cavity, where a camera is attached to the laparoscope and images of the interior patient body are

displayed onto display monitors inside the operating room. A light source is also attached to the laparoscope to illuminate the operating field, as it is essentially dark.

Before the beginning of the surgery, a Carbon Dioxide gas ( $CO_2$ ) is insufflated into the abdominal cavity. This gas separates the diaphragm wall from the interior patient organs and create a workspace for the endoscope and surgical instruments. Consequently, it introduces different organs deformations. Different endoscope types are also available, such as: flexible and capsule endoscopes that are used for examinations of the Gastrointestinal (GI) tract and commonly used for esophagoscopy, gastroscopy, and colonoscopy because of their unique abilities to reach difficult cavities.

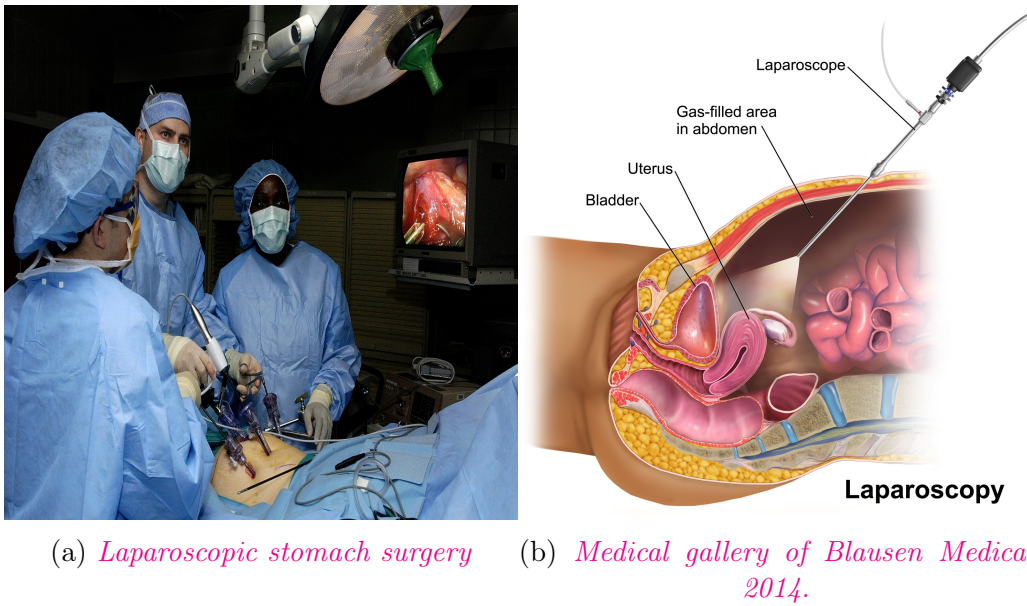


FIGURE 1.1: Typical MIS.

### 1.1.2 Medical challenges

Numerous advantages of MIS to the patient in contrast to traditional/open surgery that includes the following. Shorter hospital stays and less scarring whereas only small incisions—few millimeteres are required, this also means smaller, less noticeable scars and less bleeding. It reduces the surgical trauma and the effect on the patient immune system. Despite the patient benefits from MIS techniques, the surgeon encounters several difficulties, and thus requires skills and experience. The hand-eye coordination within a 3D scene observed on a 2D display is a major challenge. The surgeon has to overcome the natural instinct to direct the eyes to the activity of the hands, where the eye has to looks at the opposite direction of the hand motion. Secondly, the loss of binocular vision, where the surgeon watches only 2D images on the display monitor, causes visual misperceptions, mainly loss of depth perception and adds to the surgeons fatigue. This depth perception is crucial to judge the relative distance of tissues and the spatial relationship of tissues at different distances. Thirdly, the endoscope camera provides a Field Of View (FOV) that is smaller than the full FOV afforded by open surgery. More importantly, the endoscopic video does not provide information about critical anatomical

structures such as: hidden tumor and vessels underneath organ surfaces. The surgeons have to spend a large amount of extra time dissecting tissues.

### 1.1.3 Augmented reality, next step in MIS

Current medical imaging modalities such as Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI) offer the surgeons a very precise, pre-operative 3D models of different patient's organs, such as: pelvis, liver vessels, lungs, brain, heart, or bones. They enable surgeons to precisely analyze the patient's anatomy, physiology and pathology. However, these models are independent of the surgeon's viewing direction. In usual scenarios, the surgeon refers to those pre-operative models off-line on a separate display monitor, and during the surgery he/she has to mentally projects these models into the operating field.

To overcome the inherent restriction of 2D endoscopic display, on going research focus on integrating these pre-operative models with endoscopic video to guide the surgical procedure (Su et al., 2009) in the form of Augmented Reality (AR) annotations that can:

- Provide a live intra-operative identification of hidden targets (e.g. tumors, infections or foreign bodies) and critical organ structures (e.g. vessels and nerves). Thus, allowing surgeon take faster decisions.
- Free the surgeon from having to mentally match information from different sources to the scene. Thus, reducing the surgeons cognitive load during the intervention.
- Guide resections by displaying cutting trajectories and margins planned beforehand on a pre-operative model. Thus, lead to more accurate resections.
- Provide decisive information such as trocar and instruments placement, where these placements can be planned beforehand on a pre-operative model and superimposed intra-operatively onto patient body. Thus, allows for fast, safe, and optimal trocar setup to easily and effectively reach the anatomical target.
- Enhance the surgeon awareness and compensate for the limited FOV by lifting anatomical ambiguities and thus increase his/her awareness.

However, the accuracy of image augmentation with pre-operative models is very challenging and still open problem. It depends on two dynamic factors: *camera relative pose estimation* with respect to the operating field, and the *registration of these* models into the operating field. The camera relative pose tracking implies the computation, in real time, of the camera 3D location with respect to the operating field. The registration task implies anchoring, also in real time, the pre-operative models. Reaching a satisfactory accuracy in endoscopic AR is a major challenge, while maintaining it in real time is another one.

## 1.2 Our goal: dense monocular visual SLAM for MIS

This thesis focuses on the study of visual *monocular* Simultaneous Localization and Mapping (SLAM) technique in MIS to estimate two fundamental elements: *dense monocular intra-operative 3D reconstruction* of the operating field and *endoscope camera localization*. Given a sensor moving along an unknown environment, SLAM is able to simultaneously estimate both the environment structure, called map, and the sensor location with respect to that map.

### 1.2.1 Why vision-based endoscope tracking?

Traditional endoscope tracking devices, such as Polaris by Northern Digital, Inc., can provide accurate relative endoscope camera poses with respect to the operating room, however they have several limitations. They cannot provide directly the camera pose with respect to the internal surgical environment, which in most applications is needed. Secondly, the “*line of sight*” problem of the optical markers requires careful planning of the tracking devices inside the operating room and hence limits the movements of the surgeon and the assistant staff. Thirdly, they require more equipment in the operating room, and can add to setup time because a hand-eye calibration is required. Furthermore, the optical markers are not attached to scope tip, so pose uncertainty propagates significantly to the endoscope’s tip.

In contrast, vision-based tracking approaches do not have any of the limitations mentioned above, where it uses only the captured images to track the camera pose in real time. Despite vision-based solutions are challenging, but have received a particular attention over past decades and widely accepted in the clinical routine as they are easy to setup and to use inside the operating room and do not require bulky equipments. The fact that it requires the sole input of RGB images, would allow these approaches to work with any type of endoscope such as flexible ones and even capsule endoscopy.

### 1.2.2 Why dense?

When assuming accurate endoscope camera tracking is achieved, the other challenge for AR is the registration task between pre-operative models and 2D endoscopic images, which remains a difficult problem. To facilitate such registration, intra-operative reconstruction approaches have been proposed (Mountney et al., 2010a; Mirota et al., 2012; Lin et al., 2013; Grasa et al., 2014). These approaches reconstruct only image features, and they are few due to specularities, deformations and poor-texture typically exist in endoscopy, and hence these approaches are sparse and poorly describing the surgical scene. This sparse representation is not reliable for such registration, even when meshed and textured (Mountney et al., 2009; Grasa et al., 2014; Chen et al., 2018), artefacts are not avoidable.

A possible alternative to solve this registration problem is by intra-operative CT acquisition (Bano et al., 2013), which offers precise information about soft tissue morphology and structure. However, they introduce significant cost, time and design requirements for operating room. Similarly, active reconstruction techniques such as structured light (Maurice et al., 2012; Lin et al., 2015), shape-from-polarization (Martinez-Herrera et al., 2013), time-of-flight (Köhler et al., 2013) and photometric stereo (Collins et al., 2012a) can recover depth and/or surface normal information, but they require new or adapted endoscope hardware, and hence so far have had limited practical use.

In contrast, vision-based intra-operative dense reconstruction would offer a cheap, non-invasive and real-time 3D models that facilitate pre-operative/intra-operative registration. Additionally, the dense organ reconstruction facilitates the extraction of 3D features for recognition and classification applications in gastro-endoscopy, e.g. polyp classification (Mesejo et al., 2016). This dense reconstruction together with the estimated relative camera pose, can also be used to compensate for breathing motion and track tissue for laser ablation. Moreover, dense organ representation facilitates anchoring AR annotations in specific regions, where salient features can not be located.

The increasing popularity of the stereo-scope leads to a revolution in dense reconstruction, where successful and accurate techniques have been proposed (Stoyanov et al., 2010; Chang et al., 2013; Penza et al., 2016). However, the baseline of the stereo-scope cameras is relatively small and fixed, thus require close working distance, whereas uncertainty in the reconstruction can be proportionally bigger with distant organs, because they yields lower parallax which is the main cue for accurate reconstruction. While dense stereo is a practical approach it is also important to note that the majority of MIS procedures are currently performed with monocular scopes. Therefore, in this thesis, we consider only a moving monocular endoscope.

### 1.2.3 Why dense SLAM in MIS is challenging?

Recently, dense or semi-dense SLAM approaches (Newcombe et al., 2011a; Engel et al., 2014; Concha et al., 2015; Pizzoli et al., 2014; Engel et al., 2018) achieves high quality dense reconstruction results in real-time. They show the advantage of reconstruction from large number of video frames taken from very close viewpoints, where photometric-consistency is possible, with appropriate smoothness priors. They have been experimentally proven to perform robustly for indoor scenes. Despite the ground-breaking results, these approaches present some limitations when addressing endoscopy, where they require constant illumination and constant irradiance (i.e. unchanged pixel brightness) with respect to the view direction. These assumptions are not valid in endoscopy where the intensive light source that is attached to endoscope tip produces a severe illumination changes as the endoscope explores the scenes, in addition to specular reflection and organs discontinuities. Furthermore, an inadequate number of images can lead to a poorly constrained initialization for the dense SLAM optimization. It is not clear in these dense approaches how many images should be collected for dense mapping, depending on camera motion and the scene structure.

## 1.3 Contributions

In this thesis, we propose and experimentally validate one of the first intracorporeal dense monocular SLAM system that is able to cope with the above mentioned challenges. The system performs live, it is purely vision-based, provides endoscope tracking and a global and consistent dense mapping. In summary, the presented dense intra-operative reconstruction together with the relative endoscope camera pose can provide the key elements for accurate surgical guidance system and AR.

We start with the preliminary research described in Chapter 3 that has introduced a video see-through AR system based on visual SLAM (Mahmoud et al., 2017a), to visualize the intra-operative/pre-operative models using a commodity Tablet-PC with a built-in camera (video of publication <sup>1</sup>). We presented a non-invasive and interactive registration pipeline that requires minimal interaction from medical stuff, thus can be integrated easily into current clinical work flow, and accurately register the patient models onto the operating field. It performs in real-time the AR rendering, robust to occlusion, and do not require external tracking devices nor artificial landmarks on the patient skin. The proposed system can significantly contribute to early surgery planning, such as: trocar or instrument placement.

---

<sup>1</sup><https://www.youtube.com/watch?v=KNd0aXDphXM>

In Chapter 4, we have directed our study to endoscopy. Endoscope camera tracking and scene reconstruction is very challenging for monocular endoscope. We have researched the use of the state-of-the-art SLAM sparse system (ORB-SLAM Mur-Artal et al., 2015) and proved that, with a adequate tuning of different system parameters, a remarkable tracking performance (Mahmoud et al., 2017b, video of publication <sup>2</sup>). However, sparse map is good for locating the camera but very poor in describing the surgical scene. Thus, it was our contribution to extend ORB-SLAM to obtain a dense reconstruction of the surgical scene within 4.9mm of accuracy (Mahmoud et al., 2017c; Mahmoud et al., 2017d, video of the publication <sup>3</sup>). It requires an initial ORB-SLAM exploration of the abdominal cavity to acquire a set of registered keyframes used in Bundle Adjustment. After the exploration phase, the system processes the acquired keyframe in a pairwise fashion with a dense stereo algorithm run on each pair. We also demonstrated the registration of intra-operative CT models with the intra-operative dense reconstruction, in order to provide a marker-less AR(video can be found at <sup>4</sup>).

In Chapter 5, we present the first monocular SLAM system able to perform live dense mapping in in-vivo MIS scenes. The system is able to cope with the poor texture, strong illumination changes, specular reflections, surface discontinuities and small deformation typically exist in laparoscopy (video of publication <sup>5</sup>). The proposed system further extends ORB-SLAM with a novel multi-view stereo-like approach. The crux of the dense reconstruction is a variational approach that combines an illumination invariant data term and a gradient Huber norm regularizer. We provide an extensive experimental evaluation and validation on real porcine sequences of abdominal cavity exploration. We also show a comparison with other dense SLAM methods showing superior performance in terms of accuracy, density and computation time. Due to the effective selection of video frames for dense reconstruction based on parallax criteria, the proposed method can outperform the pure stereo based reconstructions. Additionally, we show a qualitative evaluation on short abdominal cavity exploratory sequence of a patient and yielding a very nice reconstruction. Furthermore, the proposed dense system show promising reconstruction results on indoor sequence from public datasets.

## 1.4 Publications

- Nader Mahmoud, Óscar G. Grasa, Stéphane A. Nicolau, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. On-patient see-through augmented reality based on visual slam. *International Journal of Computer Assisted Radiology and Surgery*, vol. 12(1), pp.1–11, 2017. **KUKA IJCARS Best Paper Award**.
- Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In *Computer-Assisted and Robotic Endoscopy (MICCAI-CARE)*, pp. 72–83, Cham, 2017. Springer International Publishing.

---

<sup>2</sup><https://www.youtube.com/watch?v=UzPjHQX5-9A>

<sup>3</sup><https://www.youtube.com/watch?v=oG54CBzqVh0&t=12s>

<sup>4</sup><https://www.youtube.com/watch?v=R17lSiIRjbM>

<sup>5</sup><https://www.youtube.com/watch?v=RJCmUY9hBSQ&feature=youtu.be>

- Nader Mahmoud, Alexandre Hostettler, Toby Collins, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Slam based quasi dense reconstruction for minimally invasive surgery scenes. *In ICRA C4 Surgical Robots: Compliant, Continuum, Cognitive, and Collaborative*, 2017. arXiv:1705.09107.
- Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Quasi-dense reconstruction from monocular laparoscopic video. In *Surgetica conference*, 2017.
- Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Live Tracking and Dense Reconstruction for Hand-held Monocular Endoscopy. *IEEE Transaction on Medical Imaging* (Submitted February 2018).



# Chapter 2

## Literature Review

### Contents

---

<b>2.1</b>	<b>Introduction</b>	<b>9</b>
<b>2.2</b>	<b>Monocular visual cue</b>	<b>10</b>
2.2.1	Sparse approaches	10
2.2.2	Dense approaches	14
<b>2.3</b>	<b>Stereo visual cue</b>	<b>18</b>
2.3.1	Only dense	18
2.3.2	Stereo SLAM	19
<b>2.4</b>	<b>Dataset</b>	<b>20</b>
<b>2.5</b>	<b>Discussion</b>	<b>21</b>

---

### 2.1 Introduction

This chapter provides a review of the related works proposed in the literature of computer vision and MIS for estimating camera pose and/or recovering scene geometry. We are interested in vision based approaches and exclude methods that use fiducial markers, where marker-based tracking methods, e.g. Mirota et al., 2009; Lapeer et al., 2008, achieve millimeter tracking accuracy, however, they can easily fail when markers are occluded, which is so common inside the operating room. For ease following the chapter, Figure 2.1 provides a broad overview of the different approaches for tracking camera motion and/or recovering scene geometry depending on the visual cue used.

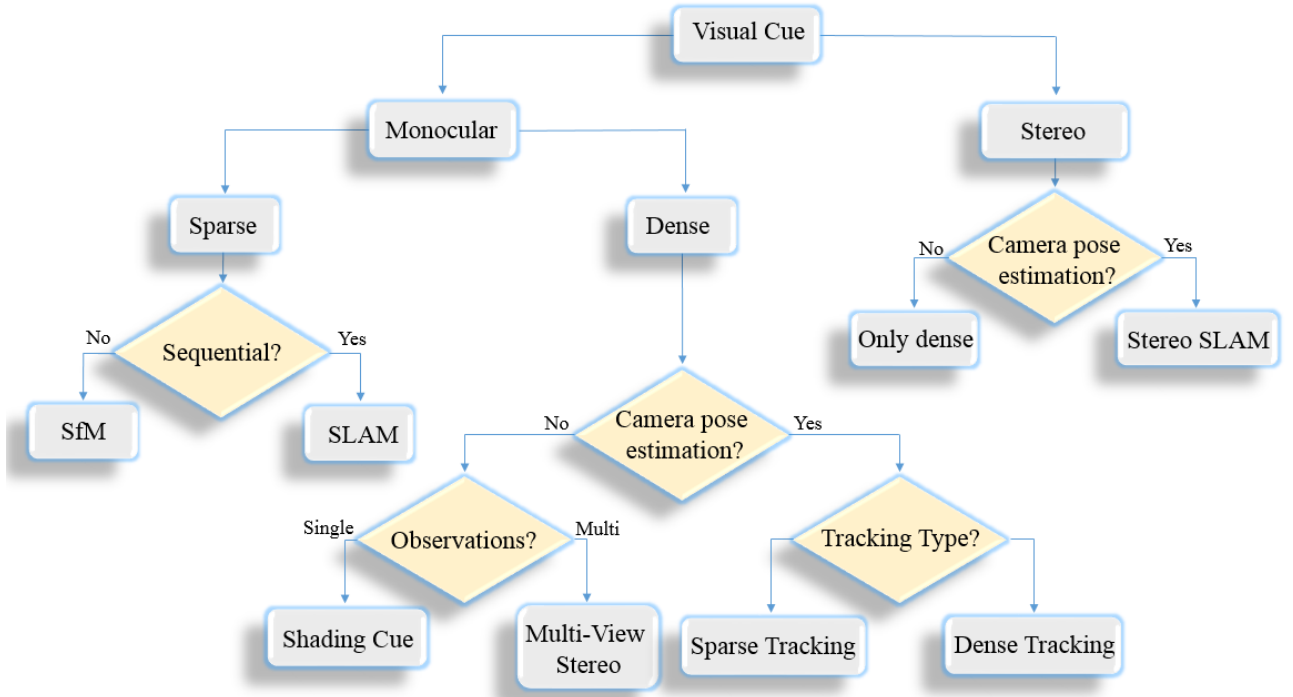


FIGURE 2.1: Taxonomy of related works.

## 2.2 Monocular visual cue

This section describes monocular approaches that effectively recover the *sparse* and *dense* 3D scene reconstruction using either single image or many images taken from different direction of the observed scene, in addition to estimate the camera pose in the given set of images.

### 2.2.1 Sparse approaches

A rich history of offline, online or recursive estimation of camera poses and sparse 3D structure begins with *Structure from Motion* (SfM) from computer vision. In essence, the estimation process involves three main stages: 1) extraction of images features (e.g., points of interest, lines, ...etc.) and matching them between images; 2) estimate camera motion from features trajectories; 3) recover the 3D structure using the estimated motion and feature trajectories. The visual systems described here produce sparse maps consists of simple geometric abstractions of scene points.

#### 2.2.1.1 Factorization based

The evolution of the SfM has noticed different important stages, where in early ages, the seminal work of Longuet-Higgins, 1981 introduced the first linear method based on point correspondences, later named the *eight point algorithm*, to solve the SfM problem for a pair of images, and estimate their relative motion and the 3D scene points. Tomasi et al., 1992 have introduced the *factorization* method for SfM assuming a simple orthographic projection camera model, in which the 3D points are measured via parallel projections onto image plane. In this work,

feature trajectories are grouped into one matrix that is later decomposed into camera motions and structure matrices using Singular Value Decomposition (SVD). The factorization methodology is later extended to include different camera projection models (Poelman et al., 1997, Triggs, 1996), imperfect data (Aanæs et al., 2002), uncalibrated cameras (Han et al., 2003), and different geometric representation (Quan et al., 1996).

The traditional factorization method has attracted the attention in MIS, where Wu et al., 2007 has employed the factorization method within a constraint-based approach for 3D reconstruction in monocular endoscopy. Wu et al., 2007 used different geometric constraints as inputs to guide the reconstruction process. Another factorization based approach was proposed by Mahmoud et al., 2012 to tackle the problem of feature occlusions and produce a photo-realistic reconstruction. The factorization approaches can provide a camera trajectory and sparse map from scratch, however they are not very accurate and require a set of well-tracked features as they are very sensitive to outliers. Nonetheless, they are good to provide an estimate of the camera trajectory and sparse reconstruction offline, however for AR purposes an online camera pose estimation is a prerequisite.

### 2.2.1.2 Bundle adjustment based

Typically, the number of 3D structure points in an SfM instance is much greater than the number of corresponding images (e.g., for an unordered collection of about  $10^3$  images, it is reasonable to expect an order of  $10^8$  scene points for a rich 3D reconstruction). The fundamental challenge for SfM is to *efficiently* process the given data, while maintaining robustness to noise/outliers in the data.

The primary class of algorithms for joint refinement of the 3D structure, the camera motion and, possibly, also the camera calibration parameters is known as *Bundle Adjustment* (BA) methods. The BA optimization is based on minimizing the reprojection error, which is the geometric error of a given 3D points cloud to match all possible observations in a set of images. Large scale BA based reconstruction (for tens of thousands of images), has been a tempting theme for computer vision researches and leads to successful and robust reconstructions from hundreds of unstructured sets of images (Snavely et al., 2006; Agarwal et al., 2010; S. et al., 2011; Crandall et al., 2011). To reconstruct the scene, an initial set of estimates for camera poses and 3D structure are required (typically, computed by a simpler techniques, and hence *factorization* is a valid option as it can compute these estimations from scratch).

The BA based approaches have encourage researches to develop more robust 3D reconstructions in MIS. Atasoy et al., 2008 has proposed an image mosaicing approach based on BA for fibroscopic video to expand the surgeon’s FOV. They used a feature based registration method to register successive endoscopic video frames, while employed bundle adjustment to maintain global consistency. Typically, a wide-baseline feature matching algorithm is required to establish correspondences between widely-separated images, thus feature occlusions and mismatches are highly likely to exist. Hu et al., 2007 and Hu et al., 2012 proposed a BA based approach to handle problems of missing feature points using a Competitive Evolutionary Agent (CEA) algorithm, and improved the robustness to outliers using trifocal tensor, that is applied in Totally Endoscopic Coronary Artery Bypass (TECAB) surgery.

As noted by Triggs et al., 2000, BA is essentially a matter of optimizing a complicated nonlinear cost function (total reprojection error) over a large parameters (3D structures and camera parameters), the accuracy of which is highly dependent on the initial estimates of the

camera poses and 3D structure. To reduce this difficulty, Sun et al., 2013 tracked the laparoscope externally to obtain a good initial camera poses for BA optimization. A major benefit of this global batch BA optimization is that the recovered structure and camera poses can achieve high accuracy. Furthermore, a video sequence is not required where feature matches between a collection of unordered and wide-baseline images are sufficient to estimate accurate reconstruction, however on the expense of longer computation times. The nature of MIS procedures is different than the general computer vision challenges that try to estimate the camera poses and scene structure from a set of images collected from the internet or databases. In MIS interventions, endoscopic video sequences are available and hence it is intuitive to use BA in a sequential manner for 3D reconstruction and camera localization. This reduces the computation complexity of the problem by optimizing a lower number of parameters, which is the backbone idea of real-time SLAM.

### 2.2.1.3 SLAM

Different from SfM, in robotics community, one main task is the real-time sensor localization, simultaneous to scene reconstruction, this problem is termed as visual SLAM. The holy grail of SLAM approaches is purely vision based, where a camera is mounted to a robot as sensor observing and moving in an unknown environment. It is a well-studied topic in robotics and more detailed survey can be found in Cadena et al., 2016. The pre-built reconstruction provide much useful information for navigation purposes, provided that localization within those maps can be performed accurately.

#### 2.2.1.3.1 Early motion estimation approaches

Burschka et al., 2005 proposed an early system to simultaneously estimate the endoscope pose and the real scaled reconstruction in sinus surgery. The real scale is recovered by a registration with a pre-operative CT model, while the camera poses are estimated in a frame basis using correspondences detected between successive frames. This approach lacks the robustness to outliers and mismatches. Thus, Wang et al., 2008 employed SIFT detection algorithm followed by SVD-based matching method between successive frames, the list of correspondences are further refined by a novel M-estimator termed Adaptive Scale Kernel Consensus (ASKC). Mirota et al., 2012 has further extended the ASKC approach and developed a robust tracking and registration system, in which an initial reconstruction is obtained from first image pair and registered to an isosurface segmented from CT images. After the initial registration, the system tracks image features and estimate camera pose with a robust 2D-3D pose estimation.

Generally, these approaches rely on feature trajectories between successive video frames for camera poses estimation and 3D reconstruction. As a result, they cannot perform well in case of large translations (i.e. changes in viewpoints) and illumination changes that are expected in endoscopy, because feature tracking is very fragile at textureless regions, and can suffer from drifts, drop-off due to occlusion, sudden camera motion, and motion blur.

### 2.2.1.3.2 Filter based SLAM

The classical age has witnessed the introduction of the probabilistic formulation for SLAM. Harris et al., 1988, introduced a single camera tracking and mapping system capable of real-time operation on very modest hardware by the use of Kalman Filter (KF). Each 3D point in the map has an associated covariance matrix that encodes its uncertainty, hence Harris et al., 1988 estimates the sequential camera motion by exploiting the uncertainty represented over the 3D point to constrain the search for correspondences in the live video stream. Chiuso et al., 2000 have emphasized the useful temporal constraints available to handle the tradeoff between the ease in solving the correspondence problem and reconstruction accuracy. Chiuso et al., 2000 have presented an early Extended Kalman Filter (EKF) based system that can handle feature occlusions and drift in scale.

Davison, 2003 has proposed a *MonoSLAM* system, establishing a milestone step towards real-time single camera visual SLAM. Aiming at accurate camera localization rather than 3D structure recovery, he allowed features to be re-used after period of occlusion, thus prevented motion drift. Davison, 2003 used a joint state representing the camera pose and the map within EKF scheme, moreover he used the joint uncertainty over predicted feature positions to reduce the computational cost of obtaining correspondences. The system's stability and agility surpassed previous systems. Montiel et al., 2006 has improved the initialization process of the *MonoSLAM* with an inverse depth parametrization of the initial map points, that enabled the representation of infinite uncertainty along the corresponding pixel ray. Since major EKF challenges are connected to efficiency and data association, Civera et al., 2009 has introduced a combination of RANSAC plus EKF, named 1-point Random Sample Consensus (RANSAC), for robust data association.

Following the inverse depth parametrization, Grasa et al., 2009 has proposed an EKF-SLAM system that perform robustly in endoscopic sequences. Grasa et al., 2009 enforced rigidity constraint and filter spurious by the use of Joint Compatibility Branch and Bound (JCBB) of Neira et al., 2001. In a latter version of the system, Grasa et al., 2011 replaced the JCBB with 1-point RANSAC detector to be able to perform in real-time, and adding a relocalization functionality, that is based on Randomized List Relocalization (RLR), to localize the camera after tracking failure. In Grasa et al., 2014, more extensive evaluation of the system was performed, more than 15 hernia repair surgeries were reported.

On one hand, it has been proven that EKF-SLAM approaches are able to perform robustly in MIS with real-time performance to estimate the endoscope camera pose, which is important step for live guidance in MIS. However, on the other hand, EKF-SLAM maps are very poor in terms of map density, it contains only very few points ranges between 10-100 points, which makes it suitable in localizing camera within very small environment. Additionally, they are known to provide less accurate results than BA approaches.

### 2.2.1.3.3 Keyframe based SLAM

The time-consuming BA has been shown to be very effective and accurate in refining scene structure and camera poses. Thus, one has to option to reduce the computational burden of the BA in robotics. The first option is to apply the BA in a hierarchical/sequential way (Hartley et al., 2003), however it does not greatly solve the computation time. Thus, it is necessary to take an alternative method whose purpose is to decrease the number of parameters to be optimized.

Shum et al., 1999 exploits information redundancy in images and divide the sequence into segments, from which local 3D reconstructions is obtained. To efficiently bundle adjust the 3D structure from all segments, they reduce the number of frames in each segment by introducing a representative frame called *virtual keyframes*, that are used during global BA. Mouragnon et al., 2006 has introduced a fast and local BA approach that ensures both good accuracy and consistency. The system significantly reduces the computational complexity compared to global BA, where a keyframe selection mechanism is developed to select only important frames, thus local BA is performed when new keyframe is acquired and involve set of neighboring frames and keyframes. The system has been experimentally evaluated in long sequence about one kilometer long.

Later, Klein et al., 2007 introduced a breakthrough Parallel Tracking And Mapping (PTAM) system, which can robustly localize the camera pose in real-time and build a 3D map of desktop-like environment consists of thousands of points. Klein et al., 2007 splits the tracking and mapping into two separate threads, run in parallel. One thread deals with the task of camera pose estimation and selection of keyframes, while the other creates and updates a 3D map of image features observed in processed video frames and perform BA. New 3D points are reconstructed inside the mapping thread, with an epipolar guided search to avoid the wait for long 2D feature tracks.

Chang et al., 2012a has researched the utilization of PTAM in Robot-Assisted Laparoscopic Prostatectomy (RALP) for tracking laparoscope poses. Benefiting from the estimated camera poses, Chang et al., 2012a proposed a photo-consistency based multi-view approach for 3D-2D image registration of segmented pre-operative CT model. However, in real endoscopic scenes, there were several robustness issues in PTAM’s performance, where features are mainly tracked with a correlation patch, thus leads to tracking failure due to poor-texture, specular and deforming regions in endoscopy.

Following the venue opened by PTAM, recently ORB-SLAM has been proposed by Mur-Artal et al., 2015. The system contains several state-of-the-art additions to obtain a robust and accurate performance in large scale environments. It includes automatic initialization, covisibility information, in addition to bag of binary words for place recognition. For large scale mapping, scale aware loop closing is used. The system uses ORB (Rublee et al., 2011) for feature description and matching in all process, what boots better performance over PTAM which uses patch search.

The clear advantage of keyframe based SLAM approaches is the real-time and high quality tracking and mapping performance they can achieve. However, these approaches are aiming at providing accurate tracking results on the expense of the sparseness of the reconstructed map. That makes the obtained map good for only locating the endoscope within the abdominal cavity, which is the first key element needed for accurate AR guidance. Due to the high dependency on image features, these approaches can perform very badly in textureless areas and with rapid camera motion, due to motion blur.

### 2.2.2 Dense approaches

In contrast to feature-based approaches, the dense approaches aim to reconstruct a rich and visually appealing 3D representation of the observed scene, where the ultimate goal is to reconstruct every single pixel in the image. In this section we are interested in dense method that

relies on RGB images as the sole sensor input, where RGB-D sensor is researched (Newcombe et al., 2011b; Bylow et al., 2013), however they cannot be applied to endoscopy.

### 2.2.2.1 Shading cue

An interesting cue in computer vision for dense reconstruction is shading, called Shape-from-Shading (SfS). SfS is very attractive to researchers, because it can provide dense 3D reconstruction from a single image. SfS techniques exploit the relationship between geometry, pixels intensities and scene illumination to recover the scene geometry. It has been appealing choice for MIS applications, specially for GI applications where textural image information tends to be scarce (Tankus et al., 2004). SfS techniques assume a completely homogenous surface, presenting no self-occlusions or shadows. Most SfS formulations assume Lambertian surface due to the mathematical simplicity of the model. Other, more sophisticated models describing rough or specular surfaces are also proposed in the literature by Ahmed et al., 2007.

SfS methods vary in their formulation depending on the model of the camera and light source used. Cameras can be modeled as either orthographic (Tsai et al., 1994), perspective (Tankus et al., 2005). On the other hand, light sources are assumed to be either infinitely far away, so that light rays will be parallel to each other (Tankus et al., 2005), or close to the surface, in which case the light source can be assumed at the optical center (Prados et al., 2006).

MIS scenes provide a niche environment for SfS, due to their suitability to the camera-light source setup usually assumed and tissue homogeneity. With the assumption of coincident camera and light source, Deguchi et al., 1996 propagate depth isocontours with an iterative scheme assuming an orthographic camera model. This model was then extended by Forster et al., 2000 to take into account the radial distortion. Wu et al., 2010 has relaxed the assumptions of SfS to deal with near point light sources that are not co-located with the camera center. Moreover, they join multiple SfS reconstructions together during orthopaedic procedures by means of a modified Iterative Closest Point (ICP) algorithm, to obtain a complete reconstruction of the observed tissue. Also, the linear approach of Tsai et al., 1994 is exploited by Karargyris et al., 2011 and Turan et al., 2017 to obtain multiple SfS reconstructions for enhanced 3D panoramic visualization during capsule endoscopy.

SfS is a strong contender for dense monocular 3D reconstruction in endoscopy because: 1) It requires no correspondences, which is a difficult task; 2) It requires only a single image; 3) The lighting conditions are highly controlled; 4) It can provide superior performance in texture-less regions, which are many. However, it is a weakly constrained problem, and real endoscopy conditions often violate its core assumptions, thus require multiple depth cues in a hybrid method, such as SfM/SfS as concluded by Collins et al., 2012b.

### 2.2.2.2 Multi-view stereo

According to Seitz et al., 2006: *The goal of multi-view stereo is to reconstruct a complete 3D object model from a collection of images taken from known camera viewpoints.* Multi-View Stereo (MVS) is an active research topic in computer vision field, where nice surveys by Seitz et al., 2006 and Furukawa et al., 2015 are exist. When assuming the estimated camera poses from sparse approaches are within bounded acceptable error, recovering scene geometry task can then be more focused on computing rich scene reconstruction.

MVS approaches tend to recover the depth of every pixel in the images taken a possibly very large set of images (Furukawa et al., 2015; Zhou et al., 2013). It can produce a highly detailed 3D reconstruction that explains the images under set of assumptions. These assumptions include: rigid Lambertian surfaces, photo-consistency (Seitz et al., 2006), known object silhouettes or shape priors (Esteban et al., 2008; Heise et al., 2015). Seitz et al., 2006 developed a six-point taxonomy that helps classifying MVS approaches according to scene representation, photo-consistency measure, visibility model, shape priors, reconstruction algorithm, and initialization requirements.

In fact, the computational complexity of estimating dense geometry with MVS has been a practical barrier to its use for real-time applications, such as computer-assisted endoscopy. Collins et al., 2013 have presented a potential usage of MVS in laparoscopy to densely reconstruct the uterus surface, but *offline* and with a predefined uterus silhouettes. There is a growing interest to import MVS methods to real-time constraint, taking into account the advantages of GPU computation power (Chang et al., 2011; Tanskanen et al., 2013). On going research work also consider handling problems of varying surface albedo and illumination (Langguth et al., 2016; Queàu et al., 2017) by incorporating SfS technique. Unlike SLAM, MVS approaches *explicitly decouple* the scene reconstruction process from camera localization, which limits its usage to, for example, build 3D scenes for virtual reality applications without being able to localize the sensor within that scene. In the next section we describe methods which combine both dense reconstruction and camera localization.

### 2.2.2.3 Feature-based tracking and dense mapping

A live dense reconstruction system must cope with increased or unknown uncertainty in the camera pose estimates. Furthermore, in the live setting the data input to the system is not fixed, hence these systems must provide a solution within a constant computational cost per frame and enable ongoing incremental reconstruction. Research efforts have been invested to produce a consistent dense representation that efficiently describe the observed scene in real-time. Methods described in this section rely on sparse SLAM as back-end for live camera pose estimation and uses that information for dense reconstruction task.

Newcombe et al., 2010 made a significant performance boost towards dense real-time SLAM and showed the advantage of reconstruction from large number of video frames taken from very close viewpoints, where photometric-consistency is possible. A base mesh was initially built from the sparse PTAM map and then iteratively polished in near real-time, by cluster of frames, through solving a variational problem with a photometric data term and a smoothness prior. However, the topology of the reconstruction was limited by the initial mesh created by the sparse points. Moreover, the photometric-consistency assumption requires a constant light source. Marcinczak et al., 2014 has improved the variational approach of Newcombe et al., 2010 to handle the challenging lighting conditions in MIS by considering the spherical color model of Mileva et al., 2007 as an illumination invariant image representation.

Graber et al., 2011 presented a dense system for depth maps fusion, which works on the set of keyframes provided by PTAM and produce their corresponding depth maps, using a dense stereo *plane-sweep* algorithm of Collins, 1996. The plane-sweep approach directly enforces the epipolar geometry between a reference image and comparison images, it is equivalent to the direct search approach proposed by Newcombe et al., 2010. For depth maps fusion, a volume based reconstruction with truncated signed distance function is used. It is only validated on

a well defined model. Wendel et al., 2012 has further extended the dense approach of Graber et al., 2011 and presented a distributed system for dense volumetric reconstruction for micro aerial vehicles. In a similar vein, Chen et al., 2018 has initially filtered the outliers in the sparse SLAM map and built the surface of the organ using Poisson Surface Reconstruction of Kazhdan et al., 2006 algorithm, after a smoothing step. This method is highly dependent on quality of the initial sparse reconstruction and the observed scene, where bad/too sparse points can lead to significant errors.

Despite the good dense performance of the these approaches, it is not clear how they would perform in real MIS procedures, where plane-sweeping approach considered by Gallup et al., 2007 targets well textured scenes. Hence it can produce good results for sufficiently textured and un-occluded surfaces, which is hard to meet in MIS scenes. Moreover, plane-sweeping assumes surface are frontal parallel and slanted surface can be correctly handled by performing multiple plane-sweeps in different direction Gallup et al., 2007, which adds to computation time. The adapted variational approach proposed by Marcinczak et al., 2014 has been validated on set of images from the video sequence, however it is hard to predict how the global reconstruction of the imaged organ would be. Furthermore, their pixel-wise data term can be very sensitive to minor transformation, both in geometry (rotation and translation) and in imaging conditions (noise and blurring).

#### 2.2.2.4 Dense tracking and mapping

These approaches estimate the camera pose by minimizing a photometric error, in contrast to reprojection error used by sparse approaches, and termed as *Direct SLAM*. This photometric error is a pixel-wise cost function that is devised and optimized as the degree of similarity between all pixels in the two images, together with a certain motion parametrization for warping either image in order to align both of them. These approaches can perform a dense (Newcombe et al., 2011a; Pizzoli et al., 2014; Concha et al., 2015) (all pixels in the image) or a semi-dense (Engel et al., 2014; Mur-Artal et al., 2015) (only high gradient areas) mapping. Typically, dense reconstruction methods (Newcombe et al., 2011a; Pizzoli et al., 2014) reconstruct high quality surfaces and require GPU acceleration due to the computational cost involved, while semi-dense approaches( Engel et al., 2014; Mur-Artal et al., 2015) recover object contours and textured surface, thus does not need GPU but rely on multi-threading optimization. Direct SLAM approaches offers:

- Rich scene understanding beyond pure geometry that is useful for object or scene recognition, navigation or AR tasks.
- Robust tracking in case of motion blur or scenes with less discriminative features.

The impressive results of theses approaches, have questioned the need for features based methods and suggest an evolution from sparse SLAM to direct SLAM. Direct SLAM approaches are very promising but still a challenge, where it require the availability of the right type of camera motion and suitable scene illumination. The photometric error used by direct approaches assumes Lambertian surfaces, no gain or exposure changes between images, and no lens artifacts like vignetting. The photometric constancy does not hold for images captured over a wide baseline or when light changes significantly. These assumptions limit its applicability

to indoor/controlled environments, where changes in illumination are very small or constant. These assumptions are violated in MIS scenes, where there are severe illumination variability as the endoscope explores the operating field, due to the high intensity light source attached to the endoscope tip.

The recent *sparse visual odometry* system proposed by Engel et al., 2018 shows how to incorporate photometric calibration and exposure information in the photometric BA. A photometric BA, where several cameras and associated depth maps are jointly optimized to minimize the photometric error, is too computationally expensive. Engel et al., 2018 showed that by operating on a sparse set of pixels per image and without smoothness priors, it is affordable to perform a sliding window photometric BA in real-time and standard CPU.

## 2.3 Stereo visual cue

Stereo-endoscope becomes ubiquitous imaging modality in abdominal interventions. It gains popularity in robotic assisted surgery, where systems such as *da Vinci* become available. A massive and valuable research works have been proposed that exploit the stereo cue for dense scene recovery. Stereo matching is the process of taking left and right images and estimating a 3D surface of the observed scene by finding matching pixels in the two images and converting their 2D positions to 3D depths. The literature of dense stereo is vast and we refer to Hirschmuller et al., 2009 and Szeliski, 2010, where a detailed explanation of different stereo matching aspects such as photo-consistency measure, local (window-based) or global optimization based methods can be found. However, in this section we are interested in stereo approaches that were successfully applied in MIS.

### 2.3.1 Only dense

In this section we explore the stereo methods that are able to compute the dense reconstruction without estimating camera pose. Stoyanov et al., 2010 has presented a robust method that perform, in near real-time, dense stereo reconstruction based on belief propagation. The method initially starts with a sparse correspondences with feature matching and subsequently propagates the 3D structure into neighboring image regions. The use of Zero-mean Normalized Cross Correlation (ZNCC) as a photometric similarity measure shows a robust performance with the challenging conditions in MIS. Röhl et al., 2012 has presented a similar correlation-based propagation approach, which depends on an adaptive version of Hybrid Recursive Matching (HRM) algorithm. Totz et al., 2014, has adapted the propagation approach of Stoyanov et al., 2010 to achieve real-time results by using coarse-to-fine pyramidal scheme with GPU parallelism. In practice, endoscopic images are very challenging, therefore mismatches are more prone to exist. Bernhardt et al., 2013 proposed different confidence criteria for assessing the quality of the stereo matching, and hence greatly reduced the percentage of the outliers in the final reconstruction. Specular highlights and lighting variation in endoscopy often lead to non-reliable matches, consequently Penza et al., 2016 considered a non-parametric image transform (Banks et al., 1997), which is robust to radiometric difference (i.e. lighting changes and noise between image pair) followed by a refinement step to polish the disparity map and fill the resultant holes.

In contrast to methods combining locally applied similarity measures with spatial or temporal constraints, other approaches consider priors to guide the reconstruction process. Schoob et al., 2013 proposed a model-based approach with the assumption that soft tissue surfaces are generally continuous and smooth. They employed the thin-plate spline model to obtain a smooth depth maps. Amir-Khalili et al., 2013 has used the segmented CT model, after being registered manually to image pair as a shape regularizer, where the depth of the each pixel is computed as a weighted depth average between stereo reconstruction and CT model. Chang et al., 2013 proposed a robust stereo approach that assume a 3D cost volume to combine local matching with global smoothness optimization. The 3D cost volume consists of  $n$  slices that have the same dimensions as the stereoscopic images, where  $n$  is the range of searching disparity. A global optimization follows the cost volume construction for achieving the smoothness of the reconstructed disparity and maintaining the discontinuity at the same time. Good surveys for classification and comparing dense stereo approaches in MIS are proposed (Stoyanov, 2012; Maier-Hein et al., 2014b; Lin et al., 2016).

The clear advantage of dense stereo approaches is their ability to recover organ deformations from a single shot, without relying on a complex computations. In addition to provide a metric reconstruction free from scale factors, monocular approaches obtain the real scaled reconstruction through a registration process with pre-operative data. However, these approaches require close working distance with target organs. High uncertainties can arise in case of distant scene areas, that render low parallax, due to the limited baseline between the stereoscope cameras ( $\approx 5\text{mm}$ ). This baseline cannot be bigger as it would require larger scope diameter. While a moving monocular endoscope can render a higher parallax, which is a crucial for accurate reconstruction.

### 2.3.2 Stereo SLAM

Recovering the dense surface geometry is a challenge, but localizing the endoscope pose with respect to the recovered surface is another one. In this section we detail works able to track also the stereo scope while estimating scene reconstruction. Mountney et al., 2006 was the first to investigate the use of EKF-SLAM for stereo scope localization and mapping in MIS, where a constant velocity and constant angular velocity model was adopted to describe the stereo endoscope motion assuming smooth camera motion and rigid scenes. Mountney et al., 2010a presented a non-rigid framework to estimate the camera motion and the deforming tissue structure, where a dynamic periodic motion model is combined with EKF-SLAM to estimate the respiration cycle of the liver. These EKF-SLAM approaches recover few feature points, represented by  $25 \times 25$  window patches, to describe the scene geometry.

To improve the limited field-of-view in MIS and obtain an attractive scene representation, Mountney et al., 2009 proposed a dynamic view expansion approach, which builds a 3D textured model from the sparse EKF-SLAM reconstruction. Similar approaches were also reported by Noonan et al., 2009 and Yip et al., 2012. Due to sparse representation of the scene, artefacts are unavoidable in the final reconstruction. Totz et al., 2011 have expanded the stereo EKF-SLAM with additional virtual features, then applied dense stereo algorithm for better describing tissue surface. In practice, EKF-SLAM approaches suffer from poor map scaling, thus the dense reconstruction were limited to smaller regions.

For large scale mapping, Lin et al., 2013 have adapted PTAM for stereo endoscope, with an extension to detect and ignore the deformable mapped points. Chen et al., 2017 follow the same vein and extend the stereo ORB-SLAM with dense stereo matching algorithm to obtain a dense reconstruction of every image pair. The depth maps are aligned in a single coordinate systems and meshed afterwards. Chang et al., 2014 and Song et al., 2018 exploited the robust depth maps obtained from each image pair and use it for stereo endoscope tracking, where Chang et al., 2014 have employed the quadrifocal relationship, while Song et al., 2018 have used the dense SURF matches between existing model and left image with RANSAC scheme to estimate the rigid translation and rotation of the stereo-scope.

## 2.4 Dataset

Accuracy evaluation is a very important step towards a thorough validation for computer assisted surgery solutions, thus it is not possible to conclude this chapter without mentioning the public datasets considered for evaluating and validating MIS solutions. It is rather important for algorithms and systems that are in a research or prototype state. Making these kinds of data freely available would significantly speed up the development for many research groups. Having common data sets does not only allow researchers to spend more time on development and less time on validation, but it also allows to benchmark algorithms against each other. Table 2.1 lists publicly available MIS datasets. Hamlyn Center Laparoscope/Endoscope Video Dataset (Mountney et al., 2010b; Stoyanov et al., 2010) provide a variety of in-vivo/ex-vivo monocular and stereo video sequences from different interventions (e.g. liver, lung, heart, colon, and spleen) with different challenges such as: rapid endoscope motion, motion blur, deformation, instrument interaction, and occlusions. It provides a CT ground truth for two stereo sequences of a silicon heart phantom, recorded with a static camera, which is suitable to evaluate and compare dense stereo approaches. Malti et al., 2012 has introduced a uterus dataset, which has been recorded by a static camera and include tissue deformation caused by instrument interaction for evaluating monocular non-rigid reconstruction. Open-CAS provides a collection of datasets that includes: short simulation liver sequences (Röhl et al., 2012), liver registration (Suwelack et al., 2014), endoscopic instrument segmentation (Maier-Hein et al., 2014a), and thyroid segmentation (Wunderling et al., 2017).

The ability to match image features between laparoscopic views is essential in many MIS applications. Puerto-Souza et al., 2013 provided the Hierarchical Multi-Affine (HMA) toolbox for feature matching, it contains image pairs with various challenging conditions, such as instrument occlusion, image blur, and organ motion. Surgical instrument are sometimes hardly to be seen in endoscopic images and thus difficult to be recognized visually, research efforts are being invested for surgical instrument tracking and surgical phase identification. For benchmarking, dataset of sequences with moving surgical instruments were made publicly available (Twinanda et al., 2016; *Surgical Robot Vision* 2016) with ground truth information about positions of the surgical instruments.

TABLE 2.1: Publicly available MIS Datasets.

Dataset	Type of Data	Scene	Camera Motion	Reconstruction GT	Endoscope tracking GT
Hamlyn center (Mountney et al., 2010b Stoyanov et al., 2010)	sequence	ex-vivo/in-vivo	static/exploratory	local area	X
Uterus (Malti et al., 2012)	sequence	in-vivo	static	X	X
Dense stereo (Maier-Hein et al., 2014b)	image-pairs	in-vitro	X	local area	X
Simulation (Röhl et al., 2012)	sequence	simulation	exploratory	Yes	Yes
Surgical instruments tracking: Maier-Hein et al., 2014a Twinanda et al., 2016 <i>Surgical Robot Vision</i>	sequence/images	in-vivo	static	X	X
Feature matching (Puerto-Souza et al., 2013)	image-pairs	in-vivo	static	X	X

As indicated in Table 2.1, the available ground truth still limited in two main aspects. Firstly, a global/complete reconstruction of the imaged organ is not available, where Mountney et al., 2010b and Maier-Hein et al., 2014b provides a reconstruction ground truth of the visible organ region in the image pairs. Secondly, it lacks the ground truth for endoscope camera pose in each frame. These two limitations are often tackled by the use of simulated data that lacks of real organs textures and often include a synthetic foreground/background separation. Such simulated data are ideal and consider a perfect scenarios that does not exist in real surgeries. Consequently, there still a strong need for a real porcine, complete, and generic dataset that allow thorough evaluation and benchmarking of different proposed solutions in MIS literature similar to traditional computer vision datasets used for dense stereo matching (Scharstein et al., 2014) and robotics datasets (Sturm et al., 2012).

For the sake of practicality, in this thesis we were interested in evaluating and validating the proposed solutions on real porcine experiment from our private dataset (see Chapter 4 for more details), which contains several in-vivo and ex-vivo exploratory sequences, together with CT surfaces ground truth of the complete abdominal cavity, that have been segmented by an expert, for reconstruction accuracy assessment. Furthermore, to avoid bias in the quantitative and qualitative evaluation of the proposed solutions, we also considered different exploratory sequences from public datasets for further validation, as we show in Chapter 5.

## 2.5 Discussion

An investigation by Strasdat et al., 2012 compared filter based vs. keyframe based sparse SLAMs in terms of computational cost of the map and trajectory accuracy. The study showed that keyframe based BA approaches, in which more features are used for tracking without joint uncertainty, leads to increased accuracy and stability over filter based systems. Generally speaking, keyframe based approaches achieved a very robust and highly accurate camera pose estimation. They globally match feature descriptors and allow for very wide baseline pose estimation (large scale tasks). However, the main limitation of these approaches is that it only exploit the visual information where features or corners can be extracted. Lack of scene textures or motion blur can make these approaches to fail or perform very poorly. In addition to the reconstructed map is a sparse set of points with little use for other tasks than camera localization. The quality of the map is heavily dependent on the matching strategy followed and

number of features in the images, which are very few in MIS due to specularities caused by the high intensive light source attached to the endoscope, organ deformation, and poor textures.

In contrast, direct SLAM approaches exploit all image pixel in dense mapping and is more robust in case of motion blur or poor-texture scene. However, these approaches, by their nature, are ill-suited for wide baseline matching. Direct SLAM approaches still rely on features to detect loops, compute the associated drifts, or relocalize the camera after tracking failure. Decent discussion and comparisons of the two methods can be found in Engel et al., 2018. We can conclude that the method with best performance really depended on the target application, and for MIS the feature-based SLAM approaches were very attractive, where it showed a robust tracking and reconstruction performance with careful selection and matching strategies of image feature, but on the expense of sparse scene representation. The impractical assumptions of direct SLAM still a barrier for its usage in MIS, thus these approaches still require a considerable improvements to tackle the challenging conditions in MIS.

The work in this thesis combine the best of feature-based SLAM and direct SLAM, where discrete scene features are used for map initialization, camera localization over long trajectories, relocation after loss of tracking, and loop closure. While the dense scene representation is obtained with two proposed approaches: pairwise dense mapping (Chapter 4) and low cost MVS dense mapping (Chapter 5). We go beyond the impractical assumptions of direct SLAM in order to handle difficulties when processing endoscopic sequences, and we came up with a system that is able to cope with illumination changes, discontinuities, repetitive textures, and small deformation caused by respiration. The proposed system has been evaluated and validated on real porcine in-vive and ex-vivo sequences.

Despite, stereo scope becomes available, there is still strong scientific interest in monocular approaches for creating 3D models from the minimal case, because it can provide insights for better understanding the visual processing. In the past, the research in the challenging monocular systems has led to significant theoretical advances that can be readily transferred to the stereo processing but not the other way round. Consequently, in this thesis, we consider only monocular endoscope as the only source of information, and we use a selection criterion for scene densification that is based on parallax rendered by a moving monocular endoscope. In comparison with *single shot* dense stereo the proposed system can render high pixel parallaxes and thus increase reconstruction accuracy.

Dense monocular SfS and MVS approaches decouple the dense reconstruction from localizing the camera pose within the operating field. Hence, make it more suitable for FOV expansion and extracting the geometric properties of the observed organ, assuming the longer processing times and scalability issues of these approaches are solved. Differently, we provide a live camera pose estimation, that is crucial for different applications such as: AR, safe navigation and instrument control during Endoscopic Skull Base Surgery (ESBS) Mirota et al., 2012.



# Chapter 3

## On-Patient See-through Augmented Reality based on visual SLAM

### Contents

<b>3.1</b>	<b>Introduction</b>	<b>23</b>
<b>3.2</b>	<b>Related work</b>	<b>25</b>
<b>3.3</b>	<b>Method overview</b>	<b>25</b>
<b>3.4</b>	<b>Method description</b>	<b>27</b>
3.4.1	SLAM architecture	27
3.4.2	Alignment of pre-operative model to SLAM map	31
3.4.3	See-through AR	33
<b>3.5</b>	<b>Experimental results</b>	<b>34</b>
3.5.1	Volunteers experiments: computation time evaluation	34
3.5.2	Accuracy evaluation	36
<b>3.6</b>	<b>Conclusion</b>	<b>39</b>

### 3.1 Introduction

Patient pre-operative models are readily available through various imaging modalities. Those models are typically displayed on a display monitor, laptop or Tablet-PC. However, the medical staff has to mentally project that information onto the patient. An AR superimposition of pre-operative models onto the patient can provide the medical staff with a kind of “X-ray vision”. Such AR visualization allows for fast, safe and optimal surgical set-up to reach the target anatomical structures. Hallet et al., 2015 exploit this AR visualization and designed the trocar placement before the actual thoracic surgery. The trocar and instrument set-up was performed on the pre-operative model before the surgery (cf. Figure 3.1(a)) and their locations are superimposed afterwards intra-operatively on a static view provided by a fixed camera (cf. Figure 3.1(b-c)). However, this technique suffers from two important drawbacks.

First, the registration of the models is performed manually and needs to be recomputed after every change of the relative camera pose with respect to the patient. Second, this kind of relative motion is difficult to avoid even using bulky fixing methods for both the camera and the patient.

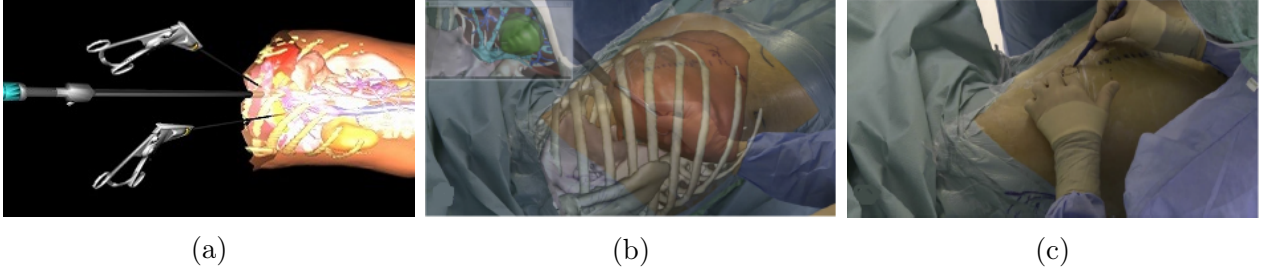


FIGURE 3.1: Port positioning with AR guidance during trans-thoracic minimally invasive hepatectomy (Hallet et al., 2015). (a) Pre-operative trocar placement planning. (b) and (c) Marking of the chosen port site.

Other successful systems provide real-time on-patient AR visualization of medical images that use: head-mounted display (Navab et al., 2007), specialized tracking hardware (Nicolau et al., 2009), or intra-operative projector systems (Sugimoto et al., 2010). However, these methods require bulky equipment such as optical tracking systems and are not appropriate for an easy integration into the operating room.

Generally, the accuracy of the image augmentation depends on two factors: 3D camera localization and model registration. On one hand, the 3D pose of the moveable camera has to be accurately estimated, in real time, with respect to the patient body. On the other hand, the pre-operative models have to be accurately registered to the patient body and have to be maintained during the camera exploration around the patient.

In this chapter, we present our SLAM-based AR approach that can accurately visualize the pre-operative/intra-operative model onto the patient skin, using only the commodity Tablet-PC with built-in camera. The sparse SLAM is used in the back-end so as to localize Tablet-PC' camera in real-time. However, the density of the obtained map was poor, it shows a good performance in localizing the camera, which is an important factor for robust AR overlay. As a front-end we proposed a non-invasive registration and visualization strategy that requires minimal interaction from medical staff. Our contributions are: 1) A SLAM for on-patient AR visualization, which only requires a Tablet-PC. 2) A usage strategy that fits the clinical constraints and is easy to setup and use inside the operating room. 3) Interactions from medical staff are reduced to the identification of 4 to 6 anatomical references at the beginning of the procedure. 4) The system is validated providing geometrical accuracy and computing cycle time. This chapter has been published in International Journal of Computer Assisted Radiology and Surgery IJCARS Mahmoud et al., 2017a.

### 3.2 Related work

Various approaches for on-patient Tablet-PC see-through AR have been proposed in recent years, that depend on two different techniques to track the camera pose: *surface-based registration* and *2D/3D point correspondences*. In surface-based registration techniques (Lee et al., 2012; Santos et al., 2014; Macedo et al., 2014; Kilgus et al., 2015; Chen et al., 2015), the Tablet-PC is mounted with a range camera, RGB-D sensor or stereo-vision to continuously capture the depth and color information, from which the skin surface is automatically extracted. This surface is then registered with the models acquired from CT images, typically ICP algorithm or its modified version is used. This process is repeated for every frame at 5-10Hz (Macedo et al., 2014). The major drawbacks of this type of techniques are the computation cost of depth image segmentation and ICP. Secondly, a good initialization for the ICP registration is required and has to be provided manually. Depending on the interface, it is not clear whether medical staff can accept this task. Thirdly, this kind of methods are not robust to surface occlusions. To achieve real-time performance, either parallel processing (Macedo et al., 2014) or client/server architecture (Chen et al., 2015) or both (Kilgus et al., 2015) are used. Where a powerful server PC is necessary to process data and the Tablet-PC is used as a display tool only.

In 2D/3D point correspondences techniques (Rassweiler et al., 2012; Müller et al., 2013; Sun et al., 2013; Schneider et al., 2014), markers need to be visible in the CT image and can be either natural landmarks or artificial ones placed on the patient before CT/MRI scanning. Their 2D positions in each frame are used to solve a 2D/3D geometrical relationship. A minimum number of markers must be visible in every frame to register the pre-operative models (Rassweiler et al., 2012; Schneider et al., 2014), which impedes the surgeon movements. Indeed, markers are likely to be occluded by surgeon hand or a surgical instruments.

To address these drawbacks, we present SLAM-based method for on-patient visualization. The utilization of SLAM for on-patient visualization purposes has been researched by Chang et al., 2012b, where marker-based tracking and SLAM are combined together. The Tablet-PC pose was tracked relative to a *specifically* designed markers, that are also used to anchor the pre-operative models. The role of the SLAM was limited to estimate the Tablet PC pose when markers are occluded. Differently, we propose a markerless Tablet-PC system that uses only the natural image feature to: track the Tablet-PC, anchor the pre-operative models, relocate the camera pose after occlusion or loss of tracking. Current on-patient visualization systems typically evaluate their accuracy by measuring the registration errors of skin fiducials as done by Kilgus et al., 2015; Santos et al., 2014; Lee et al., 2012; Chen et al., 2015, the discrepancy in pixels in the image as done by Müller et al., 2013; Schneider et al., 2014, and/or the processing time as done by Macedo et al., 2014. In contrast, our system is rigorously evaluated in terms of: processing time and robustness on human data, registration accuracy on pigs during both breath-hold and respiration phases using fiducials and registration accuracy on a liver phantom using fiducials.

### 3.3 Method overview

The overall approach is shown in Figure 3.2. It firstly consists of an offline stage. A CT volume of the patient is acquired and segmented to obtain the pre-operative models. These

pre-operative models are composed of surface meshes corresponding to the skin surface and to the selected internal body structure surfaces. The practitioner/surgeon selects at least 4 (typically between 4 and 6) anatomical landmarks ( $\mathbf{L}_i, i \in \{1..6\}$ ) on the skin mesh (called *anchor points*). The anchors should be easily identifiable on the skin of the patient during the procedure.

Inside the operating room, the practitioner directs the Tablet-PC camera at the patient and performs a translational motion to bootstrap the SLAM, once initialized he/she identify the anchor points ( $\mathbf{l}_i$ ) by clicking on the Tablet-PC live video stream. Those anchor points provide the similarity transform of the pre-operative models within the SLAM map. Afterwords, a synthetic image of the pre-operative model can be overlaid on live video stream. This interactive anchor points identification is required only once at the beginning of the procedure, then, the practitioner can move the Tablet-PC around the patient and experience the AR visualization (cf. Section 3.4.3), even if none of the anchors remains visible in the Tablet-PC camera field of view.

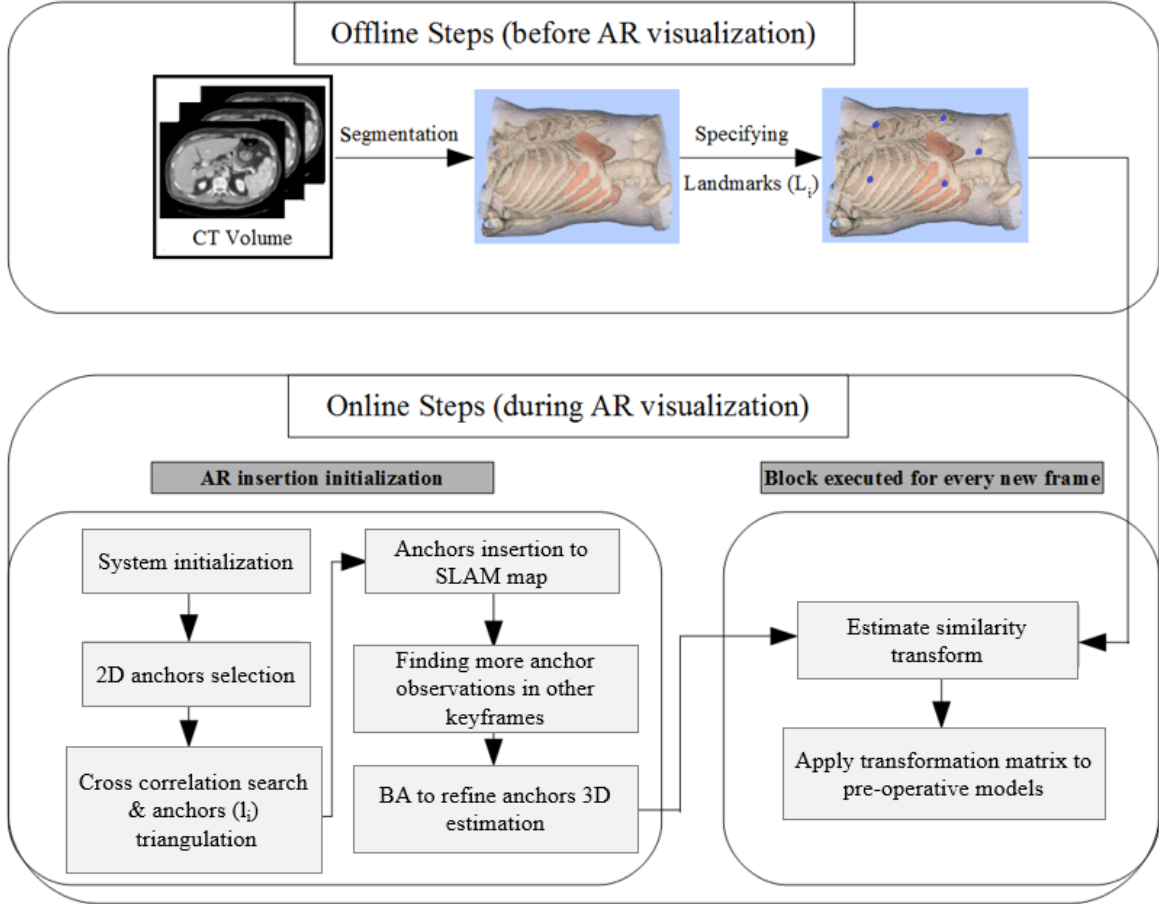


FIGURE 3.2: System workflow.

## 3.4 Method description

### 3.4.1 SLAM architecture

Our SLAM aims at running on mobile devices and is intended for small scenes. It is a PTAM-like algorithm, with two threads implementation termed *Tracking* and *Mapping* (cf. Figure 3.3), described below. The tracking thread estimates the relative Tablet-PC pose with respect to the map in real time. The mapping thread computes the 3D reconstruction of the observed scene from the live video stream. To do so, a set of interest points are matched along the sequence, we use sparse features detected in the image by the Features from Accelerated Segment Test (FAST) detector by Rosten et al., 2006. In order to reduce number of outliers, we keep only the most salient features, whose Shi-Tomasi score (Shi et al., 1994) is over 100. The ORB descriptor by Rublee et al., 2011 is then used to describe the detected features (block A in Figure 3.3). During the live camera tracking, a set of frames is selected as keyframes for non-linear BA optimization. This BA refines the estimation of the 3D map points and keyframe poses and runs in the mapping thread.

#### 3.4.1.1 Preliminaries

##### 3.4.1.1.1 Camera model

During the study in this chapter and the following chapters, we work with a pre-calibrated camera with a fixed intrinsics. The intrinsic camera matrix  $\mathbf{K}$  is computed offline from several images with a known calibration pattern. The camera calibration tool provided in *OpenCV* (Zhang, 2000) with a square calibration pattern are used to compute both  $\mathbf{K}$  and the distortion coefficient. We remove lens distortion.

$$\mathbf{K} \triangleq \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (3.1)$$

We assume a *pinhole* camera model and the *principle point* of the camera in 2D image space is represented by  $(c_x, c_y)$ , which is the center of projection or the nearest point on the image plane to the pinhole. The  $f_x$  and  $f_y$  are the horizontal and vertical focal length scaling the projected pixel in sensor width and height dimensions. In most camera sensors  $f_x \approx f_y$ , produces a square pixel. The  $s$  encodes any possible *skew* between the camera axes due to the sensor not being mounted perpendicular to the optical axis and  $(c_x, c_y)$ . The  $s$  is usually ignored and set to zero. A 3D point  $\mathbf{X}_c \in \mathbb{R}^3$  in the camera coordinate reference  $c$  is projected to a 2D image point  $\mathbf{x}$  using a projection function  $h : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ :

$$\mathbf{x} \triangleq h(\mathbf{X}_c) \triangleq \begin{bmatrix} f_x \frac{x_c}{z_c} + c_x \\ f_y \frac{y_c}{z_c} + c_y \end{bmatrix} \quad (3.2)$$

$$\mathbf{X}_c \triangleq [x_c, y_c, z_c]^T \quad (3.3)$$

$$\mathbf{x} \triangleq [u, v]^T \quad (3.4)$$

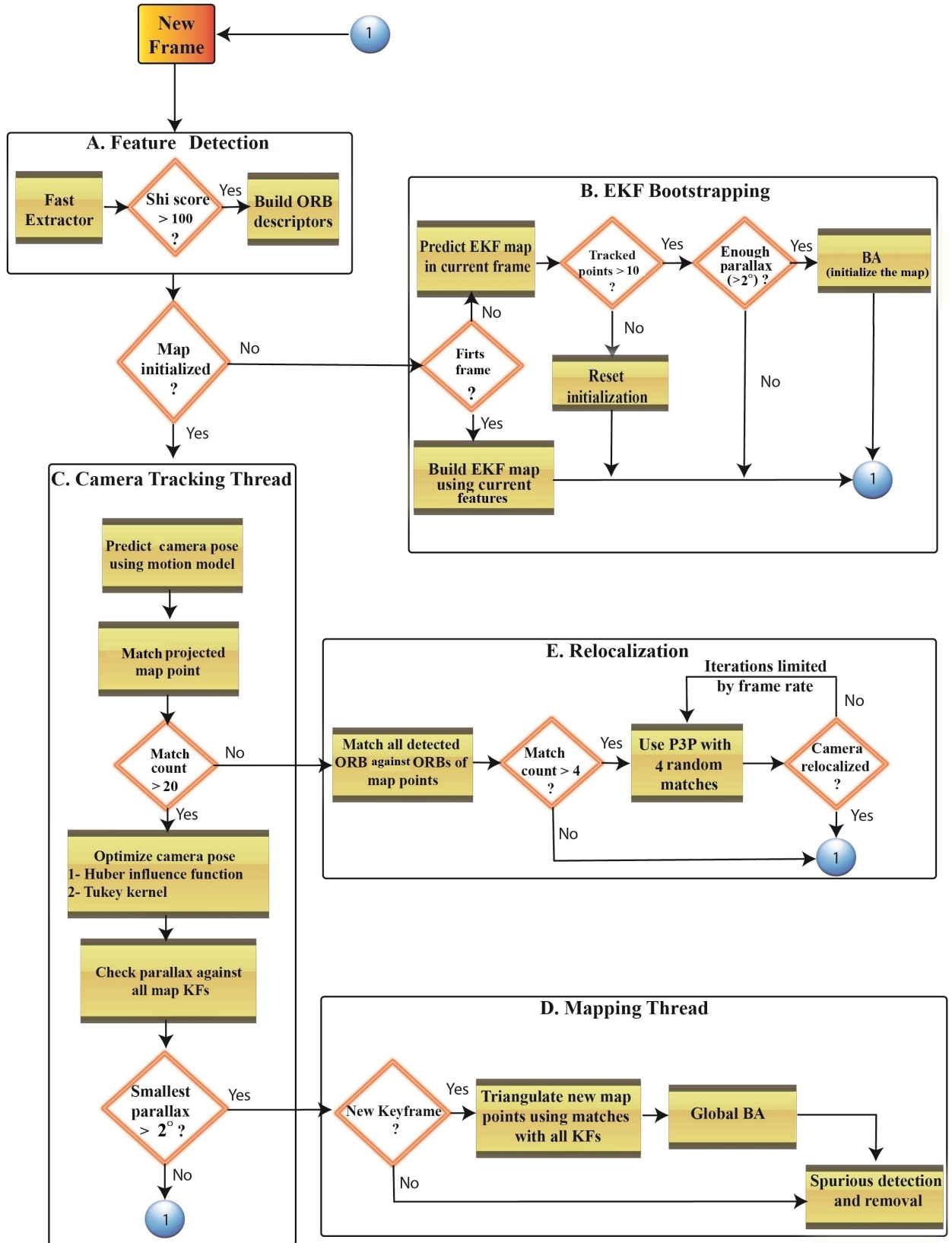


FIGURE 3.3: SLAM architecture.

Monocular cameras cannot recover the true scale of the world, and hence all the work in this chapter and the following chapters only estimate the scene map and camera trajectory up to scale.

### 3.4.1.1.2 Coordinate system

In a  $\mathbb{R}^n$  space, points are defined with respect to the *reference frame*, which consists of the basis vectors. We consider camera coordinate system has its origin at the optical center. With respect to the image, the Z axis is looking forward, the X axis is horizontal and points to the right, and the Y axis is vertical and points downwards. The 2D image coordinate has its origin at top-left corner where the  $u$ -axis points right and  $v$ -axis points down. We use the notation  $\mathbf{T}_a \in \mathbb{SE}(3)$  to denote the rotation matrix  $\mathbf{R}_a \in \mathbb{SO}(3)$  and translation vector  $\mathbf{t}_a \in \mathbb{R}^3$  of the camera at frame  $a$ .

$$\mathbf{T}_a \triangleq \begin{pmatrix} \mathbf{R}_a & \mathbf{t}_a \\ 0^T & 1 \end{pmatrix} \in R^{4 \times 4} \quad (3.5)$$

We denote world reference by  $w$ , where  $w$  represents the 3D map coordinate.  $\mathbf{T}_{c,w}$  indicates the transformation from the reconstruction coordinate  $w$  to camera coordinate  $c$ , such that a 3D point  $\mathbf{X}_w$  in the world coordinate is projected to camera frame  $c$  using:

$$\mathbf{x}_c \triangleq \pi(\mathbf{T}_{c,w}, \mathbf{X}_w) \quad (3.6)$$

$$\pi(\mathbf{T}_{c,w}, \mathbf{X}_w) \triangleq h(\mathbf{R}_{c,w} \mathbf{X}_w + \mathbf{t}_{c,w}) \quad (3.7)$$

The SLAM system computes the relative camera pose with respect to an existing map, thus for simplicity, we ignore the  $w$  and use  $\mathbf{T}_c$  to represent the relative camera pose at frame  $c$ . Rigid transformation between each local coordinate system (for each camera frame) is possible.

### 3.4.1.2 Camera tracking

This task operates sequentially over all frames of the live video (block C in Figure 3.3). A map of the scene is assumed to be available with known 3D locations of map points. On the arrival of a new frame  $i$ , the camera pose ( $\mathbf{T}_i$ ) is initially estimated using a velocity model. A correspondence search is performed to find matches of map points in frame  $i$ . Once done, this initial camera pose estimate is further optimized through a non-linear optimization of the reprojection error for the matched points. To avoid the influence of spurious matches, a two-stage optimization is applied. In the first stage the Huber influence function  $\rho_h()$  is used as it is less sensitive to outliers.

$$\underset{\mathbf{T}_i}{\operatorname{argmin}} \sum_j \rho_h(\|\mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j)\|^2) \quad (3.8)$$

where  $\mathbf{x}_{i,j}$  is the image point for  $j^{th}$  map point  $\mathbf{X}_j$  in frame  $i$ . While as a second stage the Tukey kernel is applied to achieve a robust optimization.

The difference among the various SLAM methods is how the matches between the map points and the current frame are computed. We estimate a region where the map points are expected to be found by reprojecting the 3D map points onto the predicted camera position

using eq. 3.6. The ORB descriptor of each map point is compared with those of all the features detected inside the predicted region, using the ratio between closest to second-closest neighbors as a score according to Lowe, 2004. If a match is not found, a correspondence is searched by patch correlation in the prediction region. If the number of matches is below a threshold 20 (empirically defined), the camera is assumed to be lost and the relocation process is requested.

The tracking thread also selects keyframes among the processed frames using the standard SLAM criteria: 1) minimal parallax angle with respect to all map keyframes. For each keyframe in the map, we compute the parallax with respect to the current frame using scene median depth. If the smallest parallax angle is over a threshold, the current frame is selected to be a new keyframe. This parallax threshold is set to  $2^\circ$  to control the tradeoff between the reconstruction accuracy and rapid tracking loss. The median scene point is computed using the median of the XYZ coordinates of the map points matched in the current frame.

### 3.4.1.3 Mapping

The mapping thread runs in parallel with the tracking thread but at a lower frequency, and continuously improving the estimation of the map points and keyframe poses through BA (block D in Figure 3.3). The BA minimizes the Huber robustified reprojection error of all map points with respect to all available keyframes:

$$\underset{\mathbf{T}_i, \mathbf{X}_j}{\operatorname{argmin}} \sum_{i,j} \rho_h (\|\mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j)\|^2) \quad (3.9)$$

The BA non-linear optimization is implemented using *Ceres Solver*. After each BA, a filtering step is performed to detected and remove spurious points, that is based on the following criteria. 1) the reprojection error of the point on the keyframe used for its creation is over than the median of the reprojection errors of all matched map points in that keyframe. 2) the point is matched in two keyframes only, however it is predicted (i.e. projected inside image) to be visible in more than two. 3) The ratio between number of times the point is matched and number of times is predicted is smaller than a threshold (we use 0.3). The mapping thread is also responsible for the initialization of new map points. Once a new keyframe is added to the map, matches between the new keyframe feature points and all other keyframes in the map are sought. We use standard patch correlation guided by epipolar geometry.

### 3.4.1.4 Bootstrapping

Previous mapping and tracking processes assume an existing map. Next we describe how the map is initialized from scratch (block B in Figure 3.3). For bootstrapping, the system has to select two keyframes with bigger parallax, this selection is a challenge for monocular SLAM. We use a simple EKF-SLAM with all features encoded in inverse depth (Civera et al., 2008). This approach can handle low parallax geometries and exploit every single image in the sequence to estimate the initial map and the camera poses.

We process images until most of the map points are detected with enough parallax. Then we consider the first and the last processed images as the two initial keyframes. Given these two keyframes and their relative locations, new point matches are computed by epipolar search to detect more matches when possible. Afterwards, this initial guess for the map points and two keyframe poses are fed to BA. The EKF-SLAM tries to initialize feature points that are

detected in the first frame only. If those points fail to be tracked, or go out of the field of view before obtaining enough parallax, this initial map is discarded and a new initialization is launched automatically.

#### 3.4.1.5 Camera relocation

Tracking can be lost because of camera occlusion, feature deletion due to fast camera motion or failure to track enough map points. Then it is required to re-locate the camera with respect to the existing map from scratch. We do *image-to-map* relocalization, where correspondences are sought between current image features and features in the map. Our system detects all the ORB points in the current image (block E in Figure 3.3), and hence they are matched with respect to the ORB descriptors of all the map points using as score the ratio between closest to second-closest neighbors to compute the putative matches. Then a perspective-three-point (Gao et al., 2003) from random samples of size 4 is executed. The number of RANSAC iterations is limited by the frame rate. To validate the relocation, the tracking algorithm has to produce a coherent position for the next frame in the sequence, otherwise relocation is re-attempted with the new frame.

#### 3.4.2 Alignment of pre-operative model to SLAM map

Registration is initialized interactively by the practitioner once the SLAM has been initialized. The practitioner selects the 2D anchor points over the live video stream by tapping on the tactile screen of the Tablet-PC. The 3D locations ( $\mathbf{l}_i$ ) of the 2D anchor points are computed and appended to the map following the procedure described in Algorithm 1. The pre-operative model is then translated, rotated and scaled to align the landmarks in the model with their corresponding anchors in the map (cf. Figure 3.4).

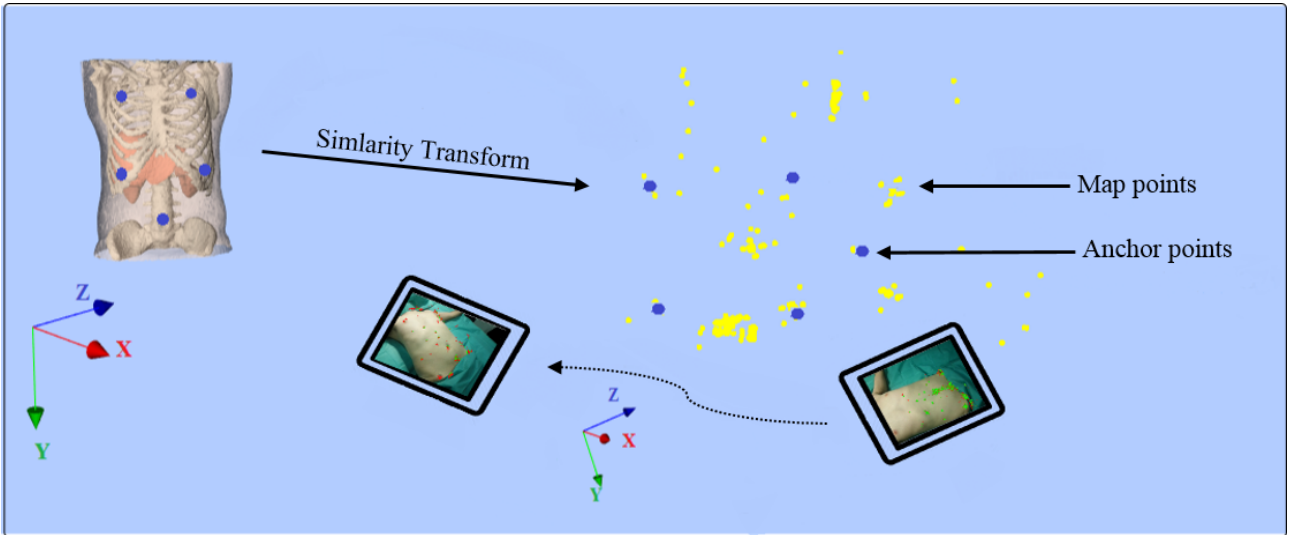


FIGURE 3.4: Pre-operative model alignment within SLAM map.

Selected anchor points store a correlation patch to match with other keyframes. The normalized correlation matching is used and guided by the epipolar geometry, and we define a

**Input** : Query image with estimated pose  $\mathbf{T}_c$  and selected 2D anchors:  $\mathbf{x}_i, i \in \{1 \dots 6\}$

**Input** : Median scene depth  $d_{scene}$  as in 3.4.1.2

**Input** : 3D landmarks from pre-operative model ( $\mathbf{L}_i, i \in \{1 \dots 6\}$ )

**Output**: Similarity transform  $\mathbf{S} \in \mathbf{Sim}(3)$  from the pre-operative model to SLAM map

**foreach**  $\mathbf{x}_i$  **do**

1) Back-project point  $i$  in query frame  $c$  coordinates:

$$\mathbf{X}_{i,c} = h^{-1}(\mathbf{x}_i),$$

$$h^{-1}(\mathbf{x}_i) = \mathbf{K}^{-1} \begin{bmatrix} u & v & 1 \end{bmatrix}^T$$

2) Define range of potential depths:

$$\mathbf{A}_{i,c} = \frac{1}{2}d_{scene}\mathbf{X}_{i,c},$$

$$\mathbf{B}_{i,c} = 2d_{scene}\mathbf{X}_{i,c}$$

**foreach** *Keyframe  $j$  with estimated pose  $\mathbf{T}_j$*  **do**

a)  $\mathbf{A}_{i,j} = \mathbf{R}_j(\mathbf{R}_c^T(\mathbf{A}_{i,c} - \mathbf{t}_c)) + \mathbf{t}_j,$

$$\mathbf{B}_{i,j} = \mathbf{R}_j(\mathbf{R}_c^T(\mathbf{B}_{i,c} - \mathbf{t}_c)) + \mathbf{t}_j$$

b) Define epipolar segment:

$$\mathbf{e}_1 = h(\mathbf{A}_{i,j}),$$

$$\mathbf{e}_2 = h(\mathbf{B}_{i,j})$$

c) Normalized cross correlation search in image segment defined by

$\mathbf{e}_1$  and  $\mathbf{e}_2$  with a bilinear interpolated warping of patch size 25x25 around  $\mathbf{x}_i$

**if** *match is found with correlation score  $> 0.8$*  **then**

    Triangulate the point ( $\mathbf{l}_i$ ) and append to the map.

    Search for more matches in other keyframes.

**else**

    continue.

**end**

**end**

**end**

BA to refine ( $\mathbf{l}_i$ ) estimations using all image observations

Estimate  $\mathbf{S} \in \mathbf{Sim}(3)$  between  $\mathbf{L}_i$  and  $\mathbf{l}_i$  that minimizes:

$$\underset{\mathbf{S} \in \mathbf{Sim}(3)}{\operatorname{argmin}} \sum_{i=1}^6 \|\mathbf{l}_i - \mathbf{S}\mathbf{L}_i\|$$

**Algorithm 1:** Pre-operative model alignment using clicked anchor points from input image.

small search segment to reduce bias in the search. Additionally, the correlation patch around point of interest is warped with bilinear interpolation to handle viewpoints changes in correlation matching. Once two keyframes with proper matches are found, the 3D location of the anchor point is triangulated ( $\mathbf{l}_i$ ), then the matches are propagated among other keyframes. The BA is executed to refine the 3D estimation of the map anchors using all their image observations. Those anchor points are never removed from the map. On the arrival of a new keyframe, we do: 1) A guided correlation for finding new anchor matches, and when found refine their 3D estimation through the BA; 2) Re-estimate the similarity transform  $\mathbf{S} \in \mathbf{Sim}(\mathbf{3})$  (rotation, translation and scaling) using Horn, 1987 to minimize the alignment error after each update in  $\mathbf{l}_i$ .

### 3.4.3 See-through AR

To provide AR overlay on the live video: 1) The SLAM tracking thread provides a pose estimate for each frame of the live stream, then a virtual camera with the same intrinsic parameters of the Tablet-PC camera is placed at the estimated pose in the virtual scene (cf. Figure 3.5(b)). 2) The image acquired by the virtual camera, taking into account the Tablet-PC camera distortion, is rendered (cf. Figure 3.5(c)). 3) The fusion is performed (cf. Figure 3.5(d)) between the real camera image (cf. Figure 3.5(a)) and the rendered one (cf. Figure 3.5(c)).

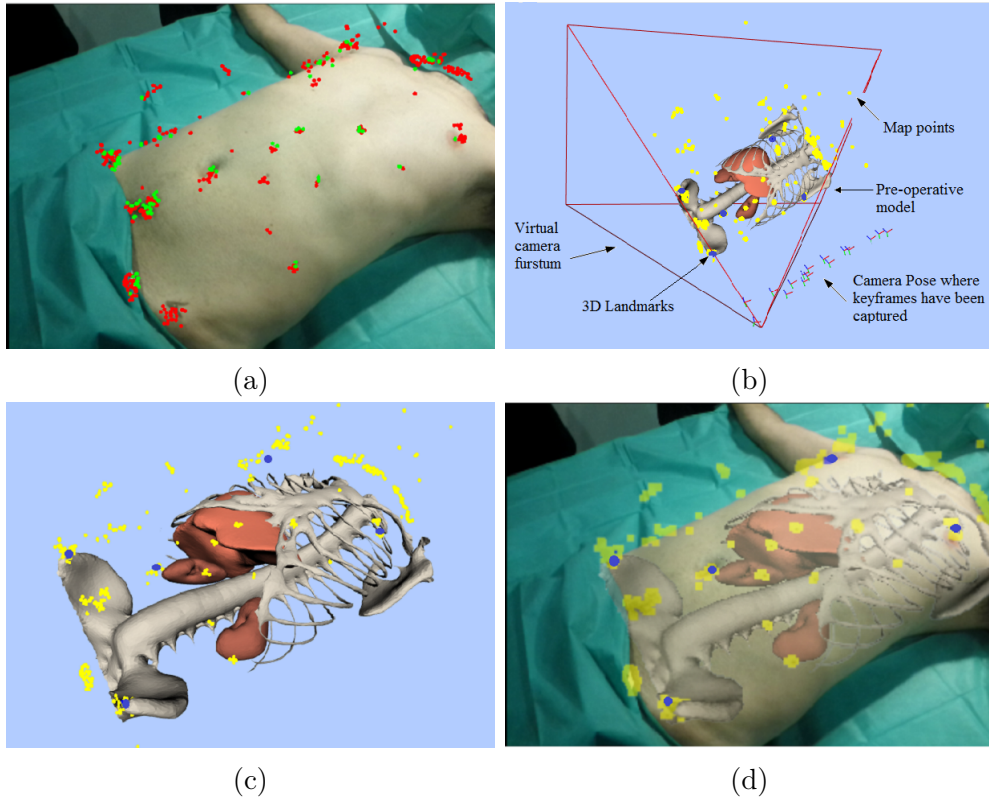


FIGURE 3.5: AR insertion. (a) Tablet-PC camera frame with projected (red) and matched (green) map points. (b) Virtual 3D scene including the registered pre-operative model. (c) Virtual camera image. (d) Fused AR image.

## 3.5 Experimental results

The proposed system has been implemented in C++ with OpenCV and VTK libraries and executed on a Sony VAIO Duo 13 Tablet-PC with Intel(R) Core i7 (1.8 GHz), 8 GB RAM, with a built-in camera of 640x480 pixels image resolution and 30 frames per seconds. Firstly, the computation time performance was evaluated in-vivo with two volunteers, each of them laying on a table while the practitioner holds the Tablet-PC and moved around them. Secondly, the system accuracy was assessed by means of several experiments with fiducials; the first experiments were on four in-vivo pigs, the second on a phantom. All computations were performed exclusively on the Tablet-PC. All pre-operative models in our experiments were segmented using our own software but can similarly be obtained using a commercial service like *Visible Patient*. More details can be seen in our video <sup>1</sup>. It is worth noting that all applicable international, national and/or institutional guidelines for the care and use of animals were followed as indicated in Appendix A

### 3.5.1 Volunteers experiments: computation time evaluation

In this experiment, the time required for each step of the system was evaluated. The CT scans of these volunteers were performed *several years ago*. For both volunteers, the five anatomical landmarks chosen as anchors for registering the pre-operative model were: right nipple, left nipple, umbilicus, right iliac crest and left iliac crest. The left and right iliac crests were marked with a pen on the skin of both volunteers, to easily identify them in the 2D images (cf. Figure 3.6(a) and Figure 3.7(a)). Figure 3.6 and Figure 3.7 show AR annotated frames for both volunteers from different points of view.

**SLAM initialization:** The SLAM bootstrapping did not fail in any of the experiments. It was initialized using on average less than 20 frames. When failed, the initialization was automatically relaunched, and eventually succeeded.

**Camera tracking and VTK rendering:** Average tracking time was approximately 32ms per frame for a map size that ranged between 180-200 points with 30-40 keyframes for video sequences composed of 750-900 frames. The average VTK rendering time was approximately 33ms per frame, including the ideal projective imaging, the distortion and the fusion with the real frame. After each mapping step, the anchors 3D locations were updated, hence the AR insertion location in the map had to be recomputed, which took less than 1.2ms. The time of initial insertion of the pre-operative model into the map can take up to 3 seconds depending on the sequence, due to searching for the anchor matches in all the keyframes. Therefore, total average time was 66.2ms per frame.

**Loss of tracking and relocalization performance:** In case of lateral (cf. Figure 3.6(f) and Figure 3.7(c)) or close up (cf. Figure 3.7(f)) Tablet-PC movements, the pre-operative models can still be registered even if most of the anchor points are not visible in the camera field of view. Camera tracking was robust to partial scene occlusion since few map points are needed for SLAM to estimate the Tablet-PC position (cf. Figure 3.10(b) and (c)). In case of full scene occlusion or severe camera motion, the relocalization module always relocated the Tablet-PC pose once a few map points were visible again, which required approximately 15ms. As a result of this module, re-initialization of the whole system in the case of tracking loss is not necessary.

---

<sup>1</sup><https://www.youtube.com/watch?v=KNd0aXDphXM>

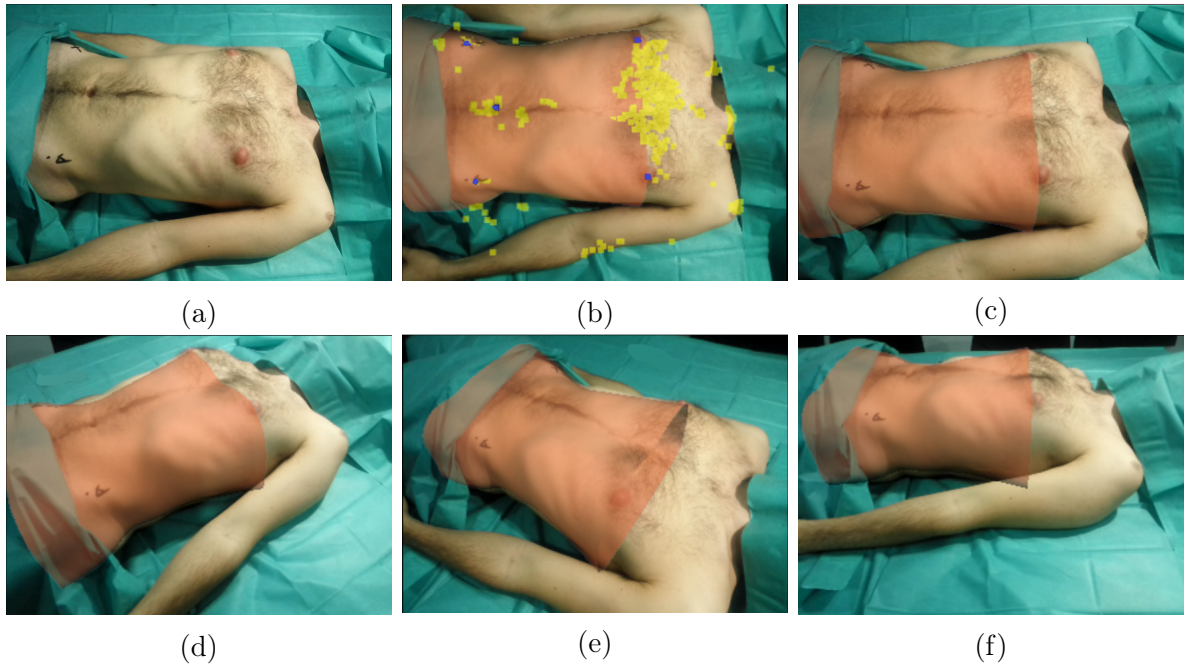


FIGURE 3.6: Experiment on first volunteer. (a) Real Tablet-PC image. (b) Skin registration with anchor points (in blue) and map points. (c-f) Skin AR overlays over frames from different points of view.

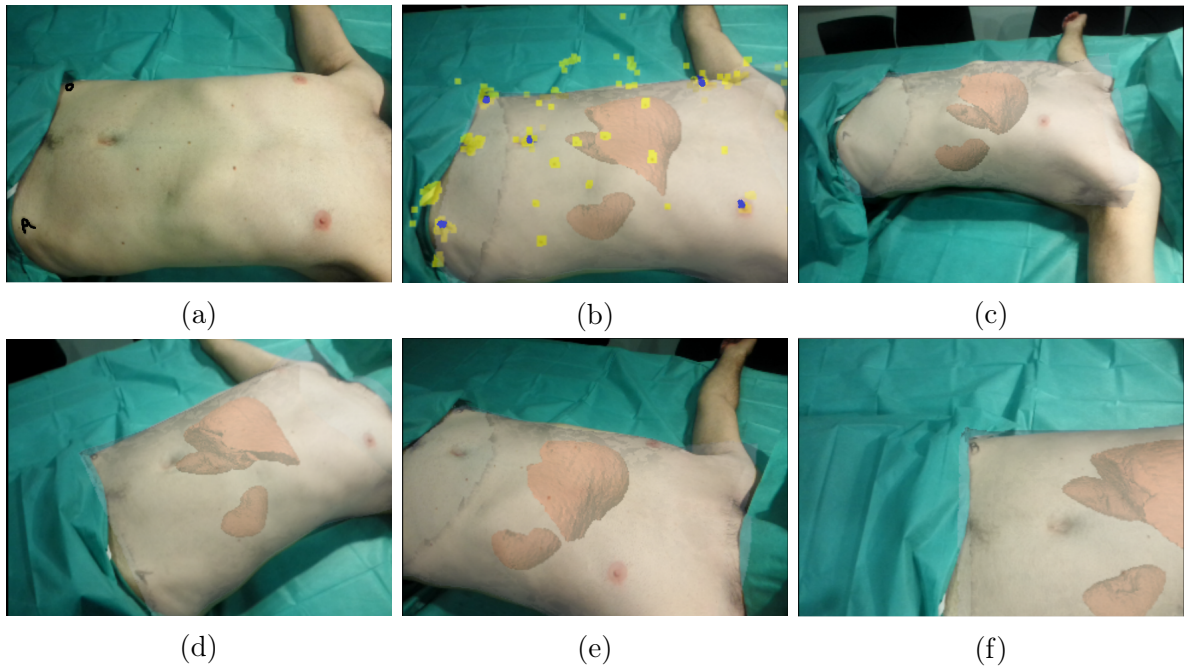


FIGURE 3.7: Registration of transparent skin, liver, left kidney and right kidney on the body of the second volunteer from different points of view.

### 3.5.2 Accuracy evaluation

To assess the registration accuracy of the proposed system, experiments on four pigs were performed and the surface Fiducial Registration Error ( $FRE$ ) as well as the Target Registration Error ( $TRE$ ) were reported. Additionally, a plastic phantom was used to evaluate the registration accuracy on internal body structures that are distant from the anchor points used for registration.

#### 3.5.2.1 Data acquisition

Each pig was placed on the CT table, and nine radio-opaque markers were stuck on its skin before the acquisition (cf. Figure 3.8(a)). The CT scan was performed during breath-hold via a mechanical ventilation system. For each pig, two videos were recorded, one with breath-hold and another during respiration. The pre-operative models and the 3D coordinates of the markers were extracted from the CT images.

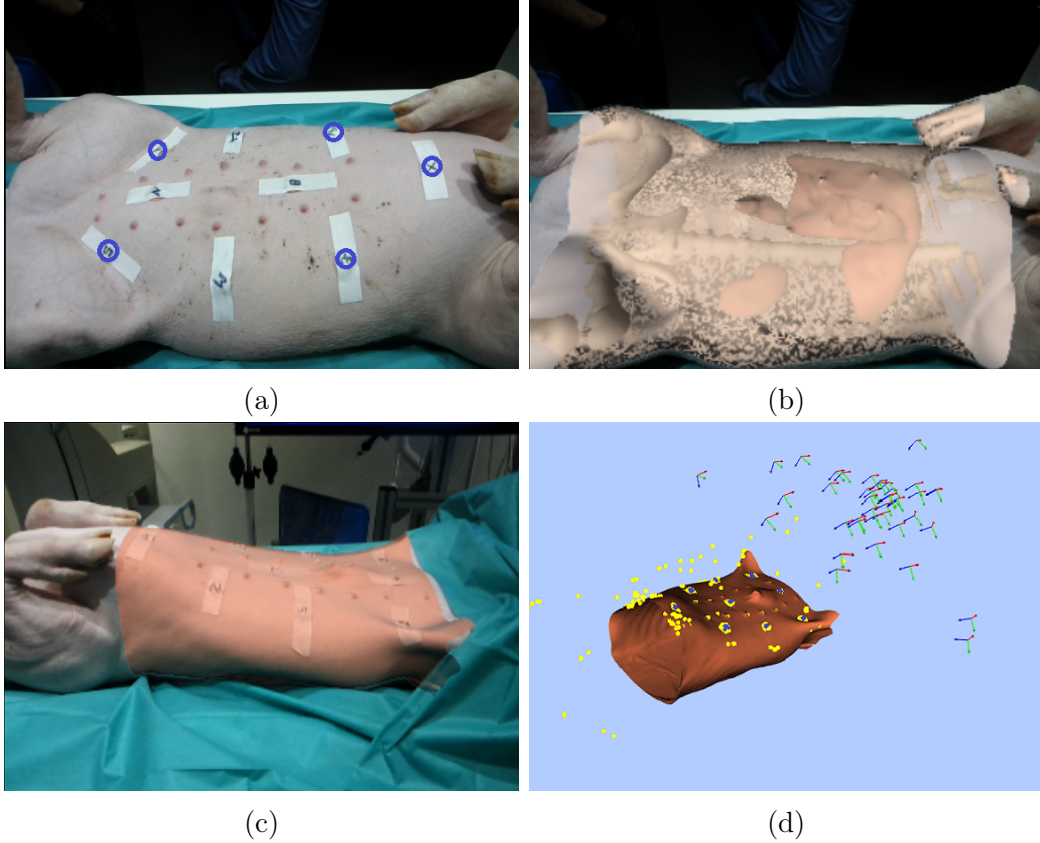


FIGURE 3.8: Experiments on pigs. (a) Nine radio-opaque markers were attached to the surface of the pig. (b) Pre-operative model composed of skin, bones, liver, left kidney and right kidney overlaid on an image of the first pig. (c) Lateral view of skin registration on the second pig. (d) Keyframe locations during camera motion around one pig.

### 3.5.2.2 Registration accuracy on pigs

The 3D coordinates of the markers extracted from CT were considered as ground truth. All markers were clicked on 2D image and their 3D estimation is computed and appended to the map. Only five markers were used as anchors to compute the 3D-3D registration those are displayed in blue in Figure 3.8(a). Figure 3.8(b) and (c) shows the registration results on two pigs from different directions. The keyframe poses during camera exploration around one of the pigs are displayed in Figure 3.8(d) and represented by axes. The averaged  $\overline{FRE}$  over all the frames in the sequence was calculated from the five markers used in the registration. The averaged  $\overline{TRE}$  over the sequence was computed from the remaining four markers.  $\overline{FRE}$  and  $\overline{TRE}$  are defined in eq. (3.10):

$$\overline{FRE} = \frac{1}{F} \sum_{f=1}^F \frac{1}{5} \sum_{i=1}^5 \|\mathbf{l}_i - \mathbf{SL}_i\| \quad \overline{TRE} = \frac{1}{F} \sum_{f=1}^F \frac{1}{4} \sum_{i=6}^9 \|\mathbf{l}_i - \mathbf{SL}_i\| \quad (3.10)$$

where  $F$  refers to the number of processed frames. As defined in eq. (3.10), the distance between the two anchor sets was computed for every frame. In the inner summation of eq. (3.10), the average distances of the five markers used for the registration and average distances of the remaining four markers were computed. Then  $\overline{FRE}$  and  $\overline{TRE}$  over all frames in the sequence were defined from the outer summation in eq. (3.10). The length of all video sequences ranged between 600 and 800 frames with 30 to 40 keyframes and map sizes between 176 and 279 points.

Each video was processed five times, each time the same frame was used to select the anchors. For the five registration trials on each pig sequence, minimum, maximum and mean values of  $\overline{FRE}$  and  $\overline{TRE}$  are reported in Table 3.1 during breath-hold. Table 3.2 shows the influence of the breathing on registration accuracy.

TABLE 3.1:  $\overline{FRE}$  and  $\overline{TRE}$  (in mm) of the four pigs sequences recorded during breath-hold.

	Pig 1			Pig 2			Pig 3			Pig 4		
	min	max	mean	min	max	mean	min	max	mean	min	max	mean
$\overline{FRE}$	2.72	3.07	2.94	2.41	2.61	2.52	3.42	4.27	3.82	1.01	2.66	1.55
$\overline{TRE}$	3.36	3.99	3.74	3.38	3.98	3.69	3.75	4.28	4.07	1.5	2.88	2.2

TABLE 3.2:  $\overline{FRE}$  and  $\overline{TRE}$  (in mm) of the four pigs sequences recorded during breathing.

	Pig 1			Pig 2			Pig 3			Pig 4		
	min	max	mean	min	max	mean	min	max	mean	min	max	mean
$\overline{FRE}$	2.53	3.64	3.11	2.76	3.7	3.1	4.62	6.09	5.32	2.62	4.22	3.1
$\overline{TRE}$	3.49	4.56	3.92	3.65	4.08	3.88	4.95	6.24	5.6	2.62	4.62	3.44

After the initial insertion into the map, the alignment of the pre-operative model is affected by a small jittering, due to the low number of keyframes at that time and thus poor geometrical conditioning. On the arrival of new keyframes providing wider baseline, thus render bigger parallax, this jittering disappears within a few seconds, according to our experiments. And hence, yielding a better estimation for the anchor points and for the similarity based alignment (cf. Figure 3.9).

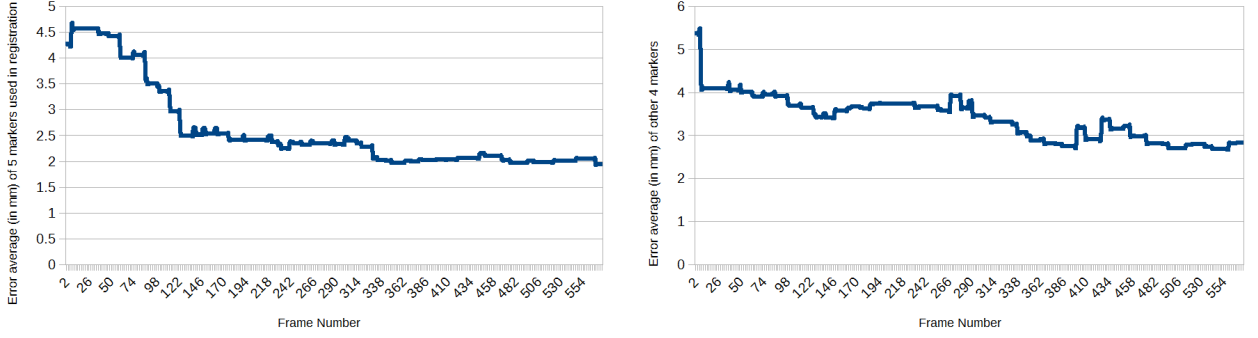


FIGURE 3.9: Evolution of the average distances error between  $\mathbf{l}_i$  and  $\mathbf{L}_i$  (in mm) over all frames of one video sequence in case of breath-hold. (a) The average distances between the 5 markers used in the registration. (b) The average distances of the remaining 4 markers.

### 3.5.2.3 Registration accuracy on phantom

In another experiment, a phantom with a *plastic liver* was used to evaluate the system accuracy for distant organ points from the body surface. 13 markers were attached on the external surface, 2 markers on the plastic liver and 4 markers on the phantom base (cf. Figure 3.10(a)). The phantom sequences and CT were obtained following the same steps as those of Section 3.5.2.1. Five markers on phantom surface were used to compute the registration (cf. Figure 3.10(a)). Table 3 shows  $\overline{FRE}$  of the five markers used for the registration,  $\overline{TRE}$  of the two liver markers and  $\overline{TRE}$  of the four markers at the phantom base. All were computed after processing the full sequence and averaging the results of five registration trials.

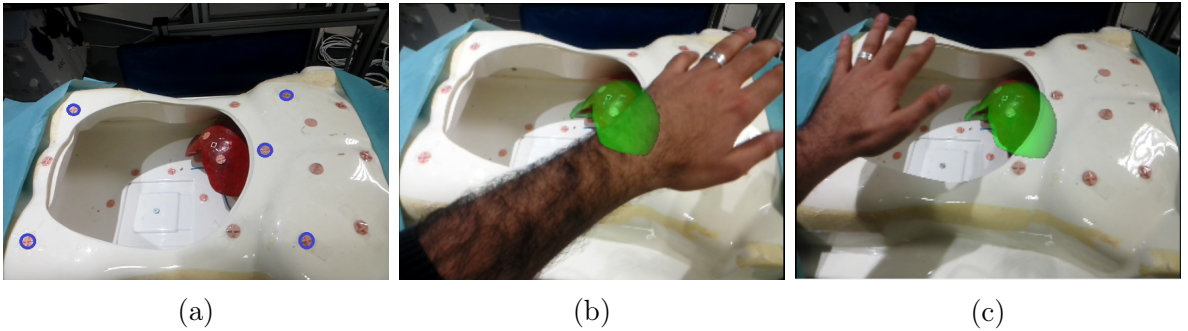


FIGURE 3.10: Experiments on phantom. (a) Markers in blue were used for the registration. (b) and (c) Liver AR overlay with partial scene occlusion.

TABLE 3.3:  $\overline{FRE}$  and  $\overline{TRE}$  (in mm) after processing the whole phantom sequence.

$\overline{FRE}$			$\overline{TRE}$ (two liver markers)			$\overline{TRE}$ (four markers at the base)		
min	max	mean	min	max	mean	min	max	mean
2.28	7.74	5.1	5.16	9.7	6.61	9.9	13.7	11.8

As can be concluded from Table 3, the closer the target to the skin, the better the registration accuracy. The four markers on the phantom base represent the worst target position, i.e. close to the skin of the back. Therefore, 11.8mm can be considered the worst system accuracy. For the sake of completeness, the registration accuracy using all the 13 markers stuck on the surface has also been computed and provide a reduction between 1.5-2.0 mm on both  $\overline{FRE}$  and  $\overline{TRE}$ . It is worth noting that the  $\overline{FRE}$  is larger than in case of pigs due to utilization of different markers. The markers used with pigs were covered and a pen was used to mark their center of mass to be easily identifiable in the images (cf. Figure 3.8(a)). In the phantom the markers were not covered and hence there were some inaccuracies in clicking the center of the cross shape of each markers.

### 3.6 Conclusion

In this chapter, we presented a real-time SLAM based on-patient see-through AR system that uses sole the input RGB image of Tablet-PC camera. In contrast to other tracking system, SLAM has proven to be able to provide a reliable camera pose estimation, even when partial scene occlusion occurs, that is important for on-patient AR tasks. We used sparse SLAM as backbone for the whole approach to provide camera pose and build a sparse map of the environment. As a front-end we proposed a registration and AR visualization strategy that requires minimal interaction from medical staff, i.e. the definition of the anchors by clicking on the live video. This interaction is considered non-disruptive by most surgeons. The proposed system provides real-time performance, robustness to occlusion and tracking failure. It can also be seamlessly integrated into the operating room as the only external device is a commercial Tablet-PC. Experimental results show the applicability of the proposed system, both in terms of computation time and accuracy. Although the system can provide a great assistance in MIS, a further extension is required. Generally speaking, internal body structures are affected by non-rigid motion during the in-hale/ex-hale cycle and thus their spatial location is changing, therefore the rigid assumption imposed in this study can produce extra error when targeting hidden anatomical structure (e.g. tumor, vessels, ...etc.). If the respiration cycle can be detected and used to update the spatial location of the pre-operatives model within the SLAM map, it can significantly reduce the error and provide a better means to reach hidden targets.

In this study, we used sparse SLAM as a robust camera estimator, which is a prerequisite for on-patient AR task. However, the quality of the obtained map was very poor with 180-200 points. The proposed AR registration strategy followed in this chapter relies on the fact that natural and anatomical landmarks can be located on the patient skin. Even when not exist, can be easily marked by the medical staff to be used afterwards in the registration. However, in endoscopy these landmarks are very few, where large smooth/poor-texture organs exist that

are dominating the endoscope FOV. Thus, anchoring pre-operative models can be a hard task and relying solely on the features is not sufficient, where feature points are not guaranteed to exist in the location of interest. Consequently, in the following chapters, we pay more attention to the quality (i.e. density) of the map and show that when properly densified, can be employed for the registration process.



# Chapter 4

## SLAM Based Quasi Dense Reconstruction For MIS

### Contents

<b>4.1</b>	<b>Introduction</b>	<b>42</b>
<b>4.2</b>	<b>ORB-SLAM overview</b>	<b>43</b>
4.2.1	Camera tracking	43
4.2.2	Mapping	44
4.2.3	Camera relocation	45
<b>4.3</b>	<b>ORB-SLAM in MIS scenes</b>	<b>45</b>
4.3.1	Parameters tunings	45
4.3.2	ORB-SLAM in action	47
<b>4.4</b>	<b>Densified discrete mapping</b>	<b>49</b>
<b>4.5</b>	<b>Quasi dense pairwise reconstruction</b>	<b>51</b>
4.5.1	Approach overview	52
4.5.2	Frame pre-processing	52
4.5.3	Building keyframe neighborhood graph	53
4.5.4	Feature based densification	53
4.5.5	Featureless depth propagation	54
4.5.6	Outliers removal and denoising	55
<b>4.6</b>	<b>Experimental evaluation</b>	<b>57</b>
4.6.1	Data acquisition	57
4.6.2	Quantitative analysis	58
4.6.3	Tracking robustness	62
4.6.4	Tuning details and computation cost	63
4.6.5	AR superimposition of intra-operative CT models	64
4.6.6	Performance on indoor sequences	65

## 4.1 Introduction

Intra-operative 3D reconstruction of the operative field from endoscope images together the relative endoscope pose are fundamental building blocks for a accurate computer-assisted guidance in MIS. However, surgical scene are challenging for vision based reconstruction because of poor/repetitive textures, occlusions, discontinuities, organ deformation and specularities caused by the high intensive light source connected to tip of endoscope. This chapter focuses on vision based endoscope camera tracking and dense pairwise 3D reconstruction of the operative field.

Mountney et al., 2006 was first to research the use of SLAM in MIS, where EKF-SLAM adapted to stereo scope and successfully applied to a *short* in-vivo sequence. Due to sparseness of the resulting map, Mountney et al., 2009 proposed to mesh and texture the obtained map to produce a photorealistic representation of the observed scene. To reduce the artefacts, Totz et al., 2011 firstly proposed to add virtual feature and exploit the availability of the stereo image pair and employ a dense stereo algorithm for better recovering the scene geometry. For long term tracking and mapping, Lin et al., 2016 has researched the utilization of PTAM for obtaining a better map and removing points created at deformable areas. These approaches exploit the stereo cue for estimating scene geometry, where robust correspondences between left and right image pairs has to be computed and then can be triangulated. In case of rectified images, this correspondence search becomes simpler and constrained in 1D search space.

Grasa et al., 2011 provided experimental evidence of the feasibility of monocular EKF-SLAM in endoscopic scenes. In Grasa et al., 2014, they provided extensive validation on in-vivo human sequences proofing its ability to be used for hernia defect measurements in hernia repair surgery. EKF-SLAM approaches have poor scaling as it is limited to smaller map sizes. Other SfM approaches were also proposed (Wu et al., 2007; Hu et al., 2012), that takes longer computation times. For real-time and high accurate camera pose estimation, external tracking sensors are considered to provide initial poses estimation (Sun et al., 2013).

Following the venue opened by PTAM in robotics community, Mur-Artal et al., 2015 proposed ORB-SLAM system, it has proven as a robust camera tracking and mapping estimator with remarkable camera relocation capabilities. In this chapter, we researched the utilization of ORB-SLAM and further extend the system to obtain a pairwise dense reconstruction in the medical arena. Our contributions are summarized as follows:

1. By careful re-tuning ORB-SLAM, we show that the endoscope pose can robustly be tracked and relocated when tracking is lost due to fast motion, image blur, or even when extracting and re-inserting the endoscope within the abdominal cavity. However, the density of the obtained map is very low, due to lack of repeatability of the features in endoscopic images.
2. We proposed a densified discrete mapping method to tackle the problem of feature repeatability, which limits ORB-SLAM from reconstructing more discrete feature points, and hence improved the estimated 3D map. Despite the improved map density, the reconstruction was limited to image areas where features are detected.

3. We proposed a pairwise dense reconstruction algorithm that significantly improves the reconstructed map density and enables reconstruction in featureless image areas. The proposed algorithm exploits the initial exploration phase, which is typically performed by the surgeon at the beginning of the surgery. We show how to convert the sparse SLAM map, with cameras accurately located, to a dense one using pairs of keyframe images and correlation-based featureless patch matching.
4. Thanks to robust endoscope tracking and dense surgical scene representation, we show a markerless AR superimposition of the liver hidden structures. The only requirement is compute the similarity transform, **Sim(3)**, that align the intra-operative models with the intra-operative dense reconstruction. Once done, the augmentation is maintained in real time while the endoscope is exploring the abdominal cavity.

The work presented in this chapter has been presented in MICCAI CARE 2016 (Mahmoud et al., 2017b), ICRA C4 Surgical Robots (Mahmoud et al., 2017c) and Surgetica conference Mahmoud et al., 2017d.

## 4.2 ORB-SLAM overview

ORB-SLAM is a PTAM-like approach and based on keyframes and nonlinear optimization. It uses multi-threading implementation for different tasks: tracking, mapping, and relocalization. It includes the covisibility information in the form of a graph as proposed in Strasdat et al., 2011. Thanks to this covisibility graph, tracking and mapping are focused on a local covisible area, independent of global environment map, which boots the system real time performance in large environment. It also includes a database of Bag Of Binary Words (DBoW2) as proposed by Galvez-López et al., 2012 for place recognition, this allows real time camera relocalization after tracking failure with significant invariance to viewpoint and illumination changes. For scale aware loop closing the method of Strasdat et al., 2010 is used. The system includes an automatic and robust initialization procedure based on model selection that permits to create an initial map of planar and non-planar scenes. The system uses FAST and ORB for feature detection and description, respectively in all tasks: tracking, mapping, relocalization and loop closing, which provide good invariance to changes in viewpoint and illumination, in addition to real-time capabilities. A complete description of the algorithm can be found in Mur-Artal et al., 2015. For the sake of completeness, we summarize next the more relevant steps: tracking, mapping, and relocation.

### 4.2.1 Camera tracking

This task tracks the camera pose sequentially in every frame of the live video stream. Assuming an existing 3D map, with ORB binary descriptors associated with map points. On the arrival of new frame, FAST features are extracted at different scale levels, and ORB descriptors are then computed. The camera pose in new frame is then estimated in three steps. In the *first step*, the pose is predicted using a constant velocity motion model, then a guided search for the map points observed in previous frame is performed. The ORB descriptor of each map point is compared with those features detected inside a search region surrounding the predicted image

location. The feature point in the search region with the smallest Hamming distance is selected as the match, only if it is over a threshold.

In the *second step*, the camera pose in current frame is optimized through a Huber robustified camera-only non-linear optimization of the reprojection error of the matched points, similar to eq. 3.8. In the *third step*, once obtained a reliable estimation of the camera pose with initial set of feature matches, the local map points defined by the covisibility graph, are then reprojected into the new frame and more correspondences are searched. A second camera-only non-linear optimization is performed considering all local map points observed in the new frame as fixed points. The tracking thread is also responsible of selecting a set of keyframes from the video sequence. These keyframes are selected based on different criteria such as: percentage of tracked points and number of frames between two neighbored keyframe.

### 4.2.2 Mapping

The mapping thread runs in parallel with the tracking thread, but at lower frequency. It is responsible of several tasks: creation of new map points, local BA, removing outliers and redundant keyframes. On the selection of new keyframe, the mapping thread computes new matches across the set of keyframes. Once the matches are available, their triangulated 3D locations are kept, only if: 1) Projection rays used to triangulate the point render parallax over a threshold; 2) Reprojection error, in the two keyframes used for its triangulation is over a threshold. The system sequentially computes new matches and iteratively improves the map accuracy through local BA that is a special case of eq. (3.9). It is performed only over a subset of neighbored keyframes  $\mathcal{K}_L$  from all available keyframes  $\mathcal{K}$  to the recently added one, this is defined by the covisibility graph. it also consider set of local map point  $\mathcal{P}_L$  seen by  $\mathcal{K}_L$ :

$$\underset{\mathbf{T}_i, \mathbf{X}_j}{\operatorname{argmin}} \sum_{i \in \mathcal{K}_L} \sum_{j \in \mathcal{P}_L} \rho_h \left( \|\mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j)\|_{\Sigma_{i,j}}^2 \right) \quad (4.1)$$

where  $\Sigma_{i,j}$  is the covariance of the position of the point  $j$  matched in keyframe  $i$ , which depends on the scale at which the feature was detected. As the camera explores new areas of the scene not imaged previously, new keyframes are added to the map, consequently new map points. Initially, map points and keyframes are initialized abundantly, then in a second stage a restrictive checks are applied to select the fittest to survive. The map point is deleted at any time when:

1. It can not be matched in at least 25% of the frames in which it is predicted to be visible.
2. It is observed in less than three keyframes.
3. It produces excessive reprojection error in the keyframes where it is observed.

These severe points selection criteria have proven essential for robust performance in endoscopic sequences. Keyframes are removed from the map when 90% of their matched map points are exist in at least three other keyframes, in order to keep just the more informative ones.

### 4.2.3 Camera relocation

Tracking can be lost because of occlusion, feature deletion due to fast camera motion, or failure to match enough map points. Therefore, the camera has to be located with respect to the map from scratch, it is known as kidnapped camera situation. To relocate the camera, ORB-SLAM detect if a query image corresponds to a revisited place using DBoW2 techniques. DBoW2 summarizes the content of the image by the visual words it contains, these visual words correspond to a discretization of the descriptor space, known as the visual vocabulary. They build on DBoW2 with ORB features, which are rotation invariant and can deal with changes in scale, so that the place recognizer can recognize places from very different viewpoints.

All the keyframes of the map are stored in a DBoW2 indexed database to recover the more similar keyframes in response to a query image. When tracking is lost, firstly, DBoW2 robustly detect a candidate keyframe of the query image from the database. This candidate keyframe contains 3D information, i.e. matched map points. Secondly, ORB correspondences are computed between the query image and keyframe candidate, and then the camera pose is estimated by using a PnP algorithm Lepetit et al., 2009 with RANSAC scheme. Once a valid camera pose is estimated, the tracking is resumed.

## 4.3 ORB-SLAM in MIS scenes

We research the utilization of ORB-SLAM for endoscope camera tracking and sparse 3D reconstruction in MIS. We first re-tune different system parameters that limit its performance in endoscopic sequences in Section 4.3.1. Secondly, we qualitatively evaluate the performance with various in-vivo sequences from different interventions (cf. Section 4.3.2).

### 4.3.1 Parameters tunings

In an initial step, we carefully re-tuned ORB-SLAM to overcome the key factors limiting its performance when processing endoscopic sequences, we report modifications relative to the ORB-SLAM standard tuning:

**Features extraction** .- The system detects best  $n$  FAST features at  $s$  image scale levels. Those features are used during tracking, mapping and relocalization tasks. We increased  $n$  by factor 2, to detect more features at  $s = 6$  levels instead of 8 with a scale factor of 1.2. We use a threshold 15 for FAST detector rather than 20.

**Initialization** .- For system initialization, features are detected in a reference frame and system tries to track those features in the subsequent frames. The search for correspondences between subsequent frames is done using descriptor matching with all ORBs detected in spatial window of size 100x100. This window size is reduced by factor 3, to avoid false positives, which are so many in endoscopy as organs have repetitive textures. The number of correspondences needed to initialize the system is also reduced by a factor of 3, where endoscopic images contains very few robust features in contrast to indoor/outdoor scenes. To recover the relative pose between first two keyframes, system uses two geometrical models, a homography for a planar scene and a fundamental matrix for non-planar scene with the normalized Direct Linear Transformation (DLT) algorithm and 8-point algorithm

respectively inside a RANSAC scheme. We increased the number of RANSAC iterations with a factor of 1.5 to reduce the influence of outliers.

**Point initialization** .- When a map point is created it has to pass different filters before added to the map, such as the following:

- **Parallax threshold:** the newly triangulated point is forced to has at least a threshold parallax to ensure that its location in 3D is accurate. This minimum parallax is increased with a factor of 5, it becomes  $1.4^\circ$ , to increase the accuracy of the triangulated points.
- **Reprojection error:** Once point is triangulated, a maximum reprojection threshold is allowed. We reduced this threshold by a factor 10, it becomes 0.6 pixel, to ensure that only rigid scene points are included eventually in the map.

**Tracking** .- During tracking process, map points are reprojected onto new image, and each one of them defines a search region for correspondence search. This matching is performed using ORB descriptors.

- **Search region:** we have increased the size of the search region with a factor 1.5. To avoid losing matches due to small deformation caused by respiration and heart beating, which are unavoidable in endoscopy.
- **Hamming distance:** we reduce the allowed Hamming distance between descriptors of matched image points by a factor 0.9. It becomes 45 bit, to enforce high similarity in the accepted matches.

**DBoW2 training dataset** .- For robust and fast relocalization, ORB-SLAM relies on DBoW2 for detecting similar keyframes to a query image. DBoW2 creates a vocabulary structured tree, in an *offline* step over a big set of descriptors, extracted from a training dataset (Bonarini et al., 2006) of indoor/outdoor environment. To build a vocabulary database dedicated to endoscopic sequences, we used a training dataset contains more than 500 images collected from Hamlyn Center Laparoscope/Endoscope Video Dataset (Mountney et al., 2010b; Stoyanov et al., 2010), *WebSurg*, and our private datasets (cf. Figure 4.1). It is worth noting that we did not notice big difference in relocalization performance, where ORB-SLAM relocates the camera pose after tracking loss with both databases. However, the time needed to load and query the new database was less than the time needed for the generic large database of ORB-SLAM. Another advantage is the lower occupancy in memory that is significant, where the size of original vocabulary was 146MB, while the new database size is 724KB.

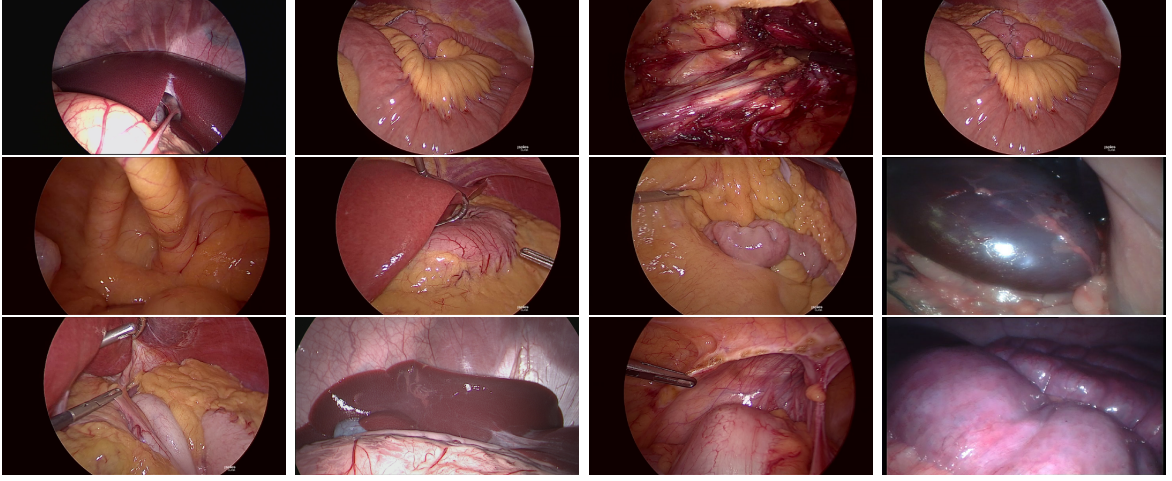


FIGURE 4.1: Image samples from training database used for DBoW2.

### 4.3.2 ORB-SLAM in action

In a second step, we qualitatively tested ORB-SLAM for laparoscope tracking and sparse scene reconstruction with different in-vivo endoscopic sequences. Figure 4.2 shows the tracking and reconstruction results of one liver sequence for a pig, where Figure 4.2(a-c) shows the input images with tracked map points, shown in green. Figure 4.2(d) shows the reconstructed map after laparoscope exploration, which consisted of 66 keyframes and 1566 map points. Figure 4.2(e-f) shows system ability to detect pig respiration, where the respiration produces a forward-backward motion of the diaphragm and hence the camera. That motion has been interpreted by the system as camera motion, where green frustum shows current laparoscope pose.

The system has also been tested with a challenging gastroscopy sequence, which contains severe reflections and abrupt movements (cf. Figure 4.3). Additionally, the system was able to robustly relocate the laparoscope camera pose after the extraction and re-insertion of the laparoscope to abdominal cavity (cf. Figure 4.4). After the exploration of the abdominal cavity has finished, the laparoscope was extracted outside the cavity while observing the liver, and is later re-inserted towards the spleen. Since several spleen points had been mapped during the exploration phase, it takes around 3 seconds to relocate the laparoscope pose.

During these sets of qualitative experiments, we concluded that the use of ORB descriptor, which is rotation and illumination invariant and also can deal with scale changes, allows for real-time and robust sparse tracking. However, there were many areas of the scene where the system was unable to track map points, being able to match only 24% of the map points visible in the image. One main reason for the matching failure, around 50% of the potential matches, is that feature detector was not able to detect repeatable points on soft organs, such as liver. Additionally, BA treats 11% of the map points as outliers (i.e. having a high re-projection error due to small deformations), this percentage raises up to 25% in areas with visually high non-rigid component. Despite the low number of matched map points, the system was able to build a sparse map and robustly track the laparoscope camera pose.

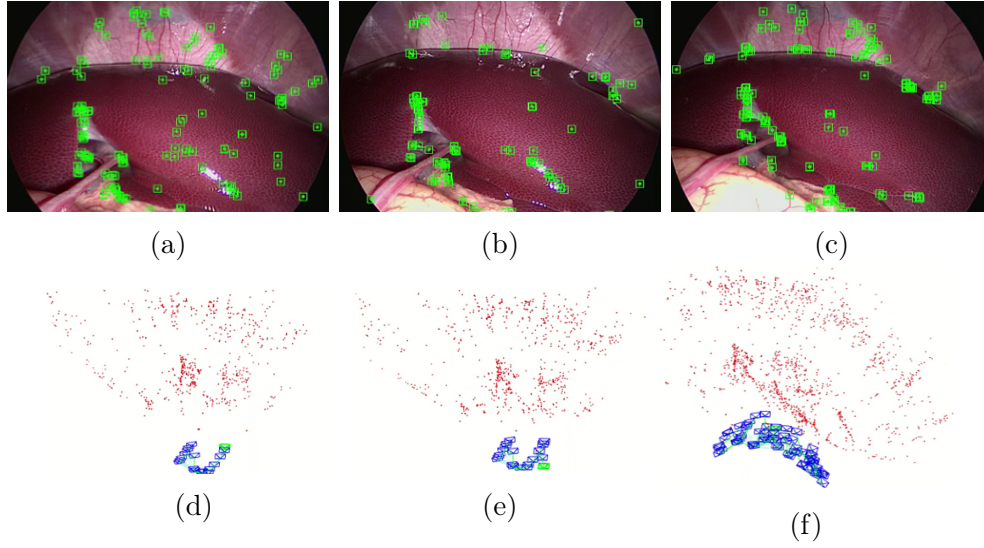


FIGURE 4.2: ORB-SLAM performance. (a-c) Image samples with projected map points in green. (d-e) Camera tracking during exploration, current endoscope pose is shown as a green frustum during inhale (d) and exhale (e), respectively. (f) Reconstructed map and keyframes locations (blue frustum), it corresponds to endoscope tip trajectory.

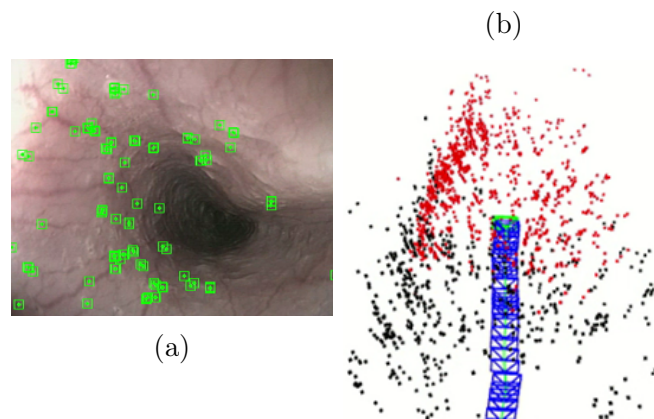


FIGURE 4.3: Gastroscopy sequence. (a) Esophagus with tracked points. (b) Reconstructed map, keyframes, and current endoscope location.

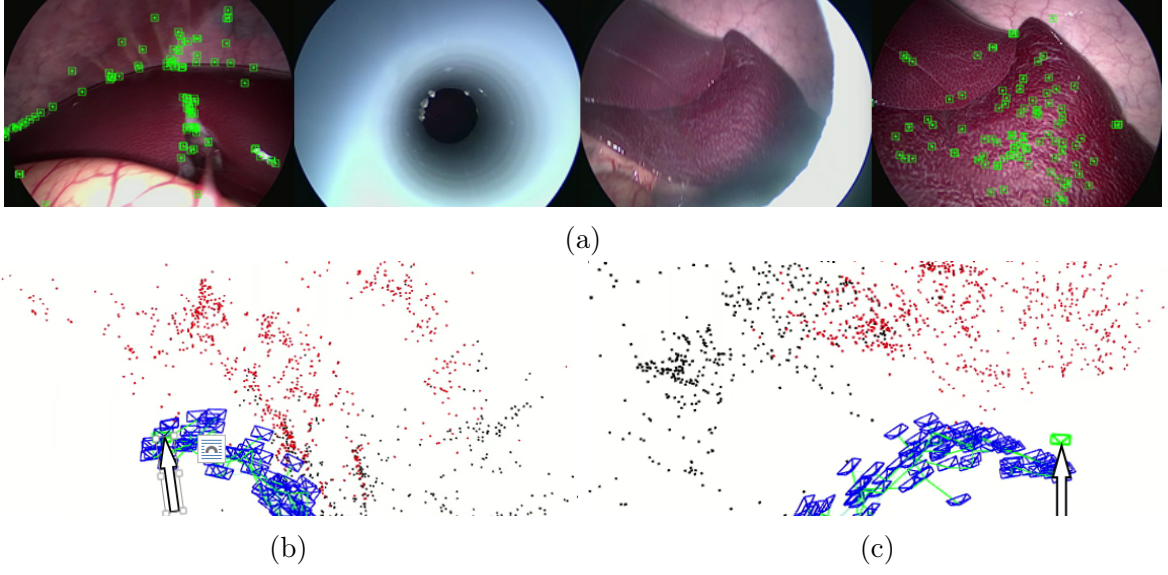


FIGURE 4.4: Relocalization. (a) Consecutive stages from left to right: successful tracking while observing the liver, tracking loss due to laparoscope extraction, laparoscope re-insertion towards the spleen, and relocating laparoscope pose and resume tracking. (b,c) The arrows refer to the laparoscope locations before and after relocalization.

## 4.4 Densified discrete mapping

The ORB-SLAM mapping thread is responsible for creating the map points and map refinement through BA. During tracking process, on the arrival of new keyframe it is sent to mapping thread and all of its feature points are matched against closest keyframe to find correspondences using descriptors search. All correspondences are then triangulated using linear triangulation algorithm to have an initial guess of their 3D locations and then appended to the map, while all unmatched features are simply ignored. One of the main challenges to feature detector algorithms in endoscopic scenes, is the repeatability to detect the same set of features across set of images. Due to lack of FAST ability to repetitively detect the same set of features during tracking, more feature correspondences cannot be found to triangulate, specially on soft organs such as liver.

In an initial extension of the ORB-SLAM system, we aimed at improving the reconstruction by increasing map density and reconstructing more discrete scene points. Next, we describe an early proposal for increasing map density. It is working in a pairwise scheme for triangulating more discrete features. On the arrival of a new keyframe  $i$  all of its FAST points  $\mathcal{F}$  are detected and their ORB descriptors are build. ORB-SLAM uses ORB descriptor to search for correspondences of  $\mathcal{F}$  in closest keyframes and all the correspondences  $f \in \mathcal{F}$  are triangulated. We extended the points initialization stage to a second step in which we try to look for correspondences for unmatched features  $f' \in \mathcal{F}$  using only one neighbor keyframe  $j$ . We use a *cross-correlation* search that is guided by epipolar geometry, with a window of 19x19, to find correspondences for  $f'$ . We do this search with only one neighbored keyframe to enable fast points initialization during live SLAM tracking. Points are initialized using this scheme during

the live tracking upon the selection of new keyframe.

To avoid correspondence search across the whole epipolar line, we constrain the search over a small segment similar to Section 3.4.2, (cf. Figure 4.5). We bound the length search segment,  $l$ , to keep computation cost low using median of all visible points in keyframe  $j$ . Two extreme points on the back-projected ray are used to bound  $l$ , which are  $P_{min}$  and  $P_{max}$ . The two points depths are computed by averaging and doubling median depth of visible map points, respectively. We then triangulate the matches and filter them for removing outliers according to three criteria:

1. Matches with correlation score less than 0.3 are eliminated.
2. Matches with negative depths are eliminated.
3. Matches with a parallax angle lower than a threshold  $1.4^\circ$  are eliminated.

The triangulated 3D location of each matched feature that passes these filters is inserted into the map.

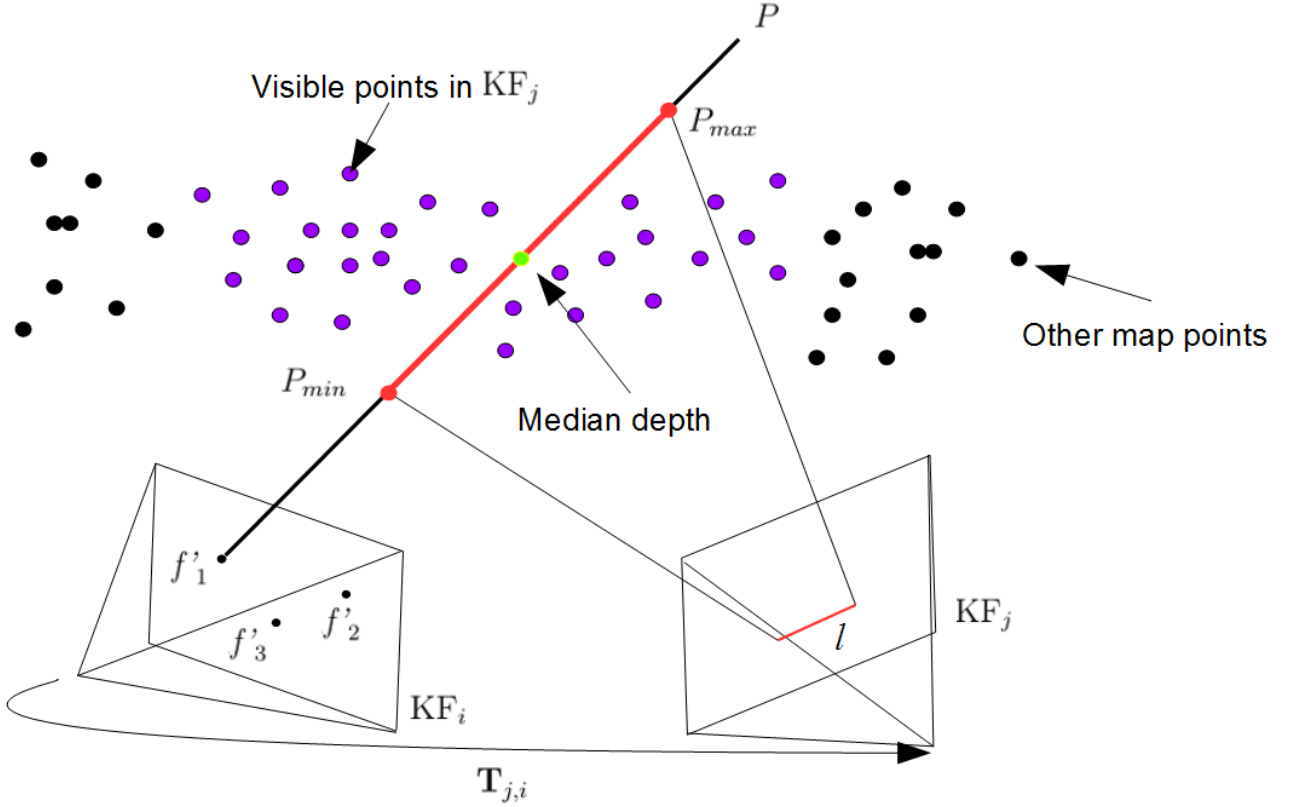


FIGURE 4.5: Epipolar guided search.  $KF_i$  is the current keyframe and  $KF_j$  is its neighbored keyframe

Figure 4.6 shows the map obtained without (Figure 4.6(b)) and with (Figure 4.6(c)) our cross correlation search for creating more discrete points. Figure 4.6(a) shows image sample with the projection of the original ORB-SLAM points in yellow. Figure 4.6(c-e) shows the incremental reconstruction during the laparoscope exploration. As can be concluded by Figure

4.6(f), a better map was obtained by reconstructing more discrete feature points in different scene areas.

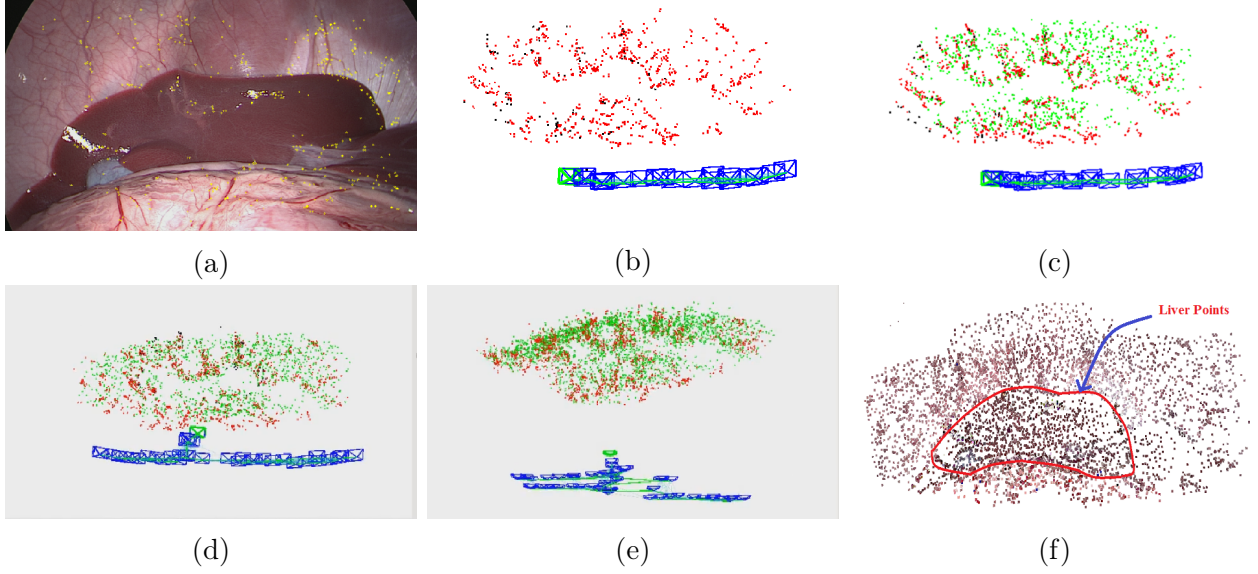


FIGURE 4.6: Improved scene mapping. (a) Image sample with original ORB-SLAM points projected in yellow. (b) Original ORB-SLAM map. (c) Map when activating cross correlation search for correspondences search. (d,e) Incremental reconstruction during laparoscope exploration. (f) Final map with points shown in original RGB intensities.

For points tracking, ORB-SLAM uses descriptor search in a predicted region in the new frame to find new matches for all map points excluding  $f'$ . For live feature tracking of  $f'$  in subsequent frames, we use Lucas-Kanade optical flow (Lucas et al., 1981). It is highly likely to loss tracking of some feature points due to image blur or fast endoscope motion, thus for those failed to be tracked we perform a guided correlation search with the patches extracted from their reference keyframe (i.e. keyframe in which points are first detected and triangulated) to estimate their 2D location in the new frame so Lucas-Kanade optical flow can re-track them in subsequent frames.  $f'$  are involved in the local BA done in the mapping thread, associated with their observations in the acquired keyframes to continuously refine/update their 3D estimation. We exclude  $f'$  from the camera-only non-linear optimization that performed during camera pose estimation to keep computation cost of this non-linear optimization low.

## 4.5 Quasi dense pairwise reconstruction

Despite the significant improvement in the reconstructed scene map in previous section, it constraints the reconstruction only to feature locations, which are few in endoscopic images and hence the reconstructed map is a discrete set of points. The sparseness of the resultant map prevents its utilization for other tasks than locating the endoscope pose within the surgical scene.

In this section, we present a quasi dense reconstruction algorithm that is able to densely reconstruct the operating environment. It accurately densify the sparse reconstructed map

computed during the exploration phase by ORB-SLAM, using pairs of keyframe images and ZNCC featureless patch matching. Only a small number of relevant keyframe pairs are selected for scene densification. They are selected using their respective baseline in the covisibility graph, and are treated as a stereo pair. Densification is then done in three main steps: *initial feature based densification* where we do a 3D reconstruction of unmatched image features (cf. Section 4.5.4), *depth propagation* where we propagate the reconstruction to featureless regions (cf. Section 4.5.5) and finally *reconstruction post-processing*, where outliers are removed and the reconstruction is smoothed (cf. Section 4.5.6). The initial feature based densification step is similar in its nature to method presented in Section 4.4 with further important improvements.

### 4.5.1 Approach overview

We outline the overall approach in Figure 4.7. During both SLAM and dense reconstruction we pre-process each frame to handle particular challenges of laparoscopic image data as explained in Section 4.5.2. During exploration, SLAM is run until the end of this phase, which typically lasts no more than a minute. The SLAM process is denoted by the top loop in Figure 4.7. This outputs a set of keyframes, their respective camera poses, a set of features detected in each keyframe and sparse 3D map. Next the three stages: feature based densification, depth propagation and reconstruction post-processing are run. Once finished, the laparoscope pose can be tracked in the incoming laparoscopic frames in real-time using the sparse ORB-SLAM map.

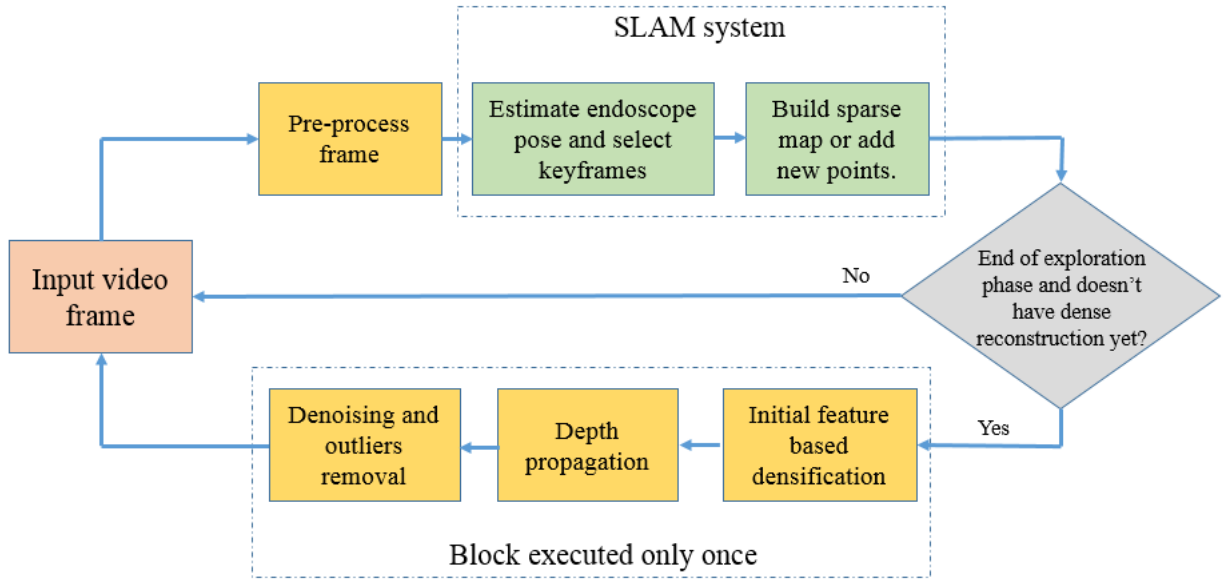


FIGURE 4.7: Quasi dense pairwise reconstruction pipeline.

### 4.5.2 Frame pre-processing

We first detect and eliminate specular reflections to avoid introducing false features to SLAM system. This is done by converting the RGB frame to HSV and thresholding the saturation component. All detected features in these areas are ignored. Most feature detectors (including

FAST) work on monochrome frames. We compute these by converting the RGB frame to monochrome using the average of the green and blue channels. This is because they give the highest contrast for human tissue according to Tromberg et al., 2000.

### 4.5.3 Building keyframe neighborhood graph

We construct a neighborhood graph  $G$  that connects pairs of SLAM keyframes. This is a sparse graph with typically  $O(g)$  edges, where  $g$  is the number of keyframes. Sparseness is necessary to keep processing time low. Each edge  $(i, j)$  in  $G$  corresponds to a stereo pair, and we use this for both feature-based densification and depth propagation. This is constructed as follows. For each keyframe  $i$  we compute the rendered parallax  $\alpha$  with respect to all neighbored keyframes as a ratio between respective baseline and map median depth. We add an edge  $(i, j)$  if  $\alpha$  falls within the range  $\alpha_1 \leq \alpha \leq \alpha_2$ . The lower-bound  $\alpha_1$  ensures there is sufficient baseline with which to reliably reconstruct points in 3D. The upper-bound  $\alpha_2$  ensures that the keyframes are not too far from each other. We give the default values of  $\alpha_1$  and  $\alpha_2$  (and all other parameters defaults) in Table 4.3.

### 4.5.4 Feature based densification

We process each keyframe pair  $(i, j) \in G$  as follows. As mentioned in Section 4.4, we have two types of features detected in keyframe  $i$ : matched  $f$  and unmatched features  $f'$ . The matched features are those that have been already matched by the SLAM system in other keyframes, have been triangulated, and have been inserted into the map. The goal of this step is to reconstruct each feature in  $f'$ . The process is similar to method described in Section 4.4 but with important improvements such as :

- Benefiting from constructed graph in Section 4.5.3, we keep searching for matches of  $f'$  in all the keyframes in graph  $G$ , not only in one neighbored keyframe. Doing this ensure obtaining higher percentage of matches.
- To gain robustness to severe illumination variability in endoscopy, we use ZNCC with window size  $\mathcal{W}$ , rather than the simple correlation patch.
- To gain robustness to small deformation typically exist in endoscopic sequences caused by respiration and heart beating, we do the correspondence search in a margin of 10 pixels around the defined epipolar segment rather than constraining the search in the epipolar segment  $l$ .
- This step is done in an offline manner after the endoscope exploration is finished.

The matches are then triangulated and filtered to remove outliers when: ZNCC score less than a threshold  $\psi$ , having negative depths, and the parallax angle lower than  $\alpha_1/2$ . The triangulated 3D position of each matched feature that passes this filtering step is then inserted into the map. Figure 4.8(b,e) shows the sparse SLAM map from the initial exploration, where map points are shown in red and keyframes poses are the blue rectangles. Figure 4.8(c,f)) shows the reconstruction after the featureless densification stage on two different sequences from our

private dataset (cf. Figure 4.8(a)) and Hamlyn Centre Laparoscopic/Endoscopic Dataset (cf. Figure 4.8(d))

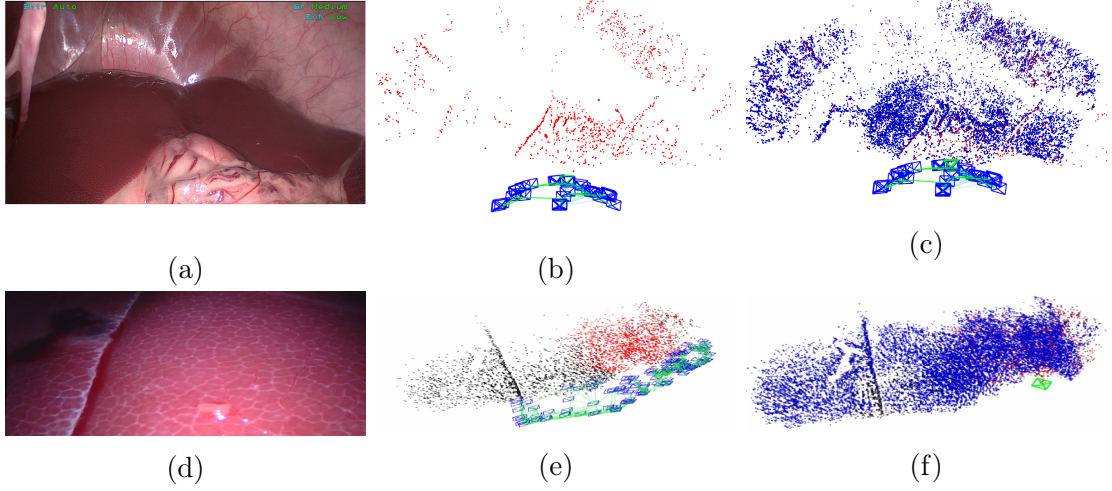


FIGURE 4.8: Feature based densification.

#### 4.5.5 Featureless depth propagation

After feature-based densification, we further densify the map at featureless regions through a depth propagation algorithm. The process works on each keyframe pair  $(i, j) \in G$  as follows. First we take all points that were matched in keyframes  $i$  and  $j$ , and their depths are used as seed depths, which are then propagated to neighboring pixels. Thus, it is important to have as more feature correspondence as possible. We then continue to propagate depth around seeds on best-first basis by popping a seeds queue, as proposed by Stoyanov et al., 2010. New matches are added to the queue as the algorithm iterates until no more matches can be popped.

Consider a seed point with a 2D position  $\mathbf{x}$  in keyframe  $i$  and  $\mathbf{x}'$  in keyframe  $j$ , with  $N(\mathbf{x})$  and  $N(\mathbf{x}')$  spatial neighbored pixels, respectively in a  $6 \times 6$  window. These seed matches are used as temporal smoothing prior to control the smoothness of the disparity estimation of all  $N(\mathbf{x})$  and  $N(\mathbf{x}')$  pixels. For each neighbored pixel  $\mathbf{m} \in N(\mathbf{x})$  a ZNCC is used to find a corresponding match  $\mathbf{m}'$  in a  $6 \times 6$  window centered in the corresponding spatial location in keyframe  $j$ , that has higher ZNCC score than  $\psi - 0.1$  and satisfy the smoothness constrain defined in eq. (4.2). We use  $\beta$  to control the smoothness of the disparity estimation. ZNCC is also used as a similarity measure during this propagation step. Figure 4.9 shows the depth propagation step on two keyframes (cf. Figure 4.9(a-c)) and on SLAM map (cf. Figure 4.9(d,e))

$$N(\mathbf{x}, \mathbf{x}') = \{(\mathbf{m}, \mathbf{m}'), \|\mathbf{m} - \mathbf{m}' - (\mathbf{x} - \mathbf{x}')\| \leq \beta\} \quad (4.2)$$

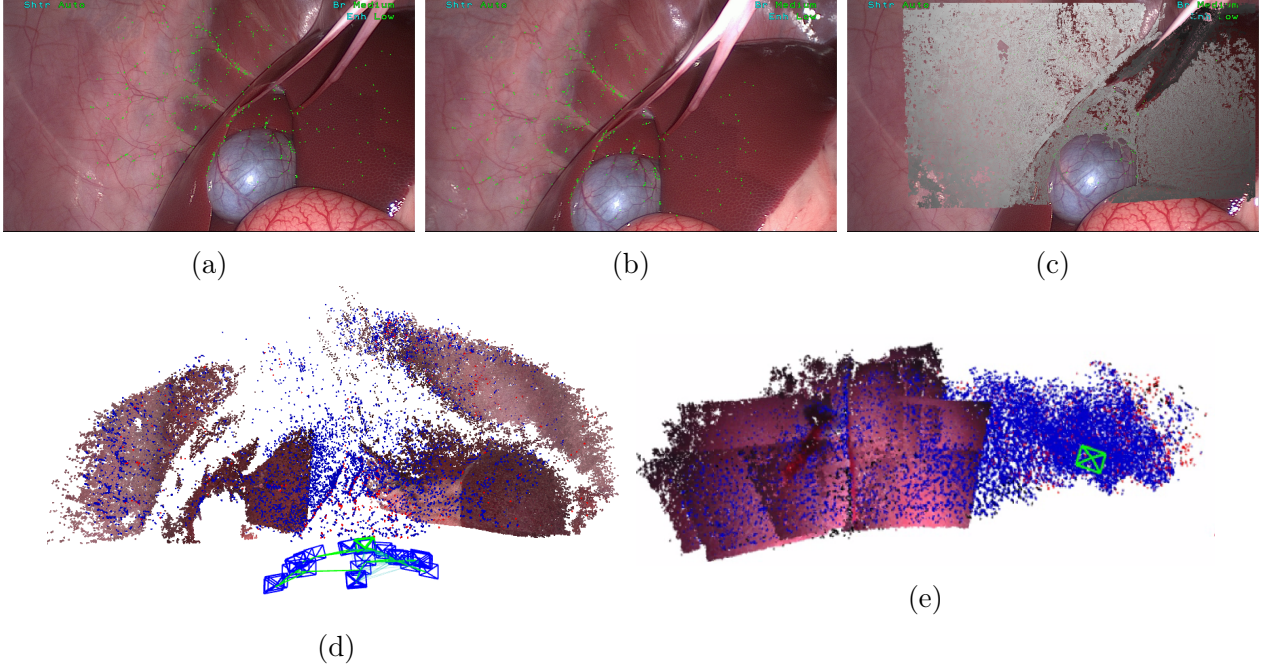


FIGURE 4.9: Featureless depth propagation. (a,b) Two keyframes with corresponding features. (c) Disparity map, where darker pixels are closer. (d,e) Depth propagation in SLAM map.

#### 4.5.6 Outliers removal and denoising

Because depth propagation operates on keyframe pairs, there will be some disagreement due to noise across different keyframe pairs, typically at very low-textured regions. We deal with this by a robust averaging and merging. First we detect any remaining outliers in the dense map using point neighborhood statistics algorithm of Rusu et al., 2008. This works by eliminating points if they are unusually far apart from the nearest  $\kappa$  points, according to a threshold  $\tau$  multiplied by standard deviation of all points distances. We then remove noise in the surface using Moving Least Square (MLS) algorithm of Alexa et al., 2003, that define the smooth surface locally in two steps. Consider a map point  $\mathbf{P}$  in Figure 4.10(a), thus in a first step, a local plane  $H_{\mathbf{P}}$  to  $\mathbf{P}$  is defined:

$$H_{\mathbf{P}} \triangleq \langle \mathbf{Q}, \mathbf{n} \rangle + D \quad (4.3)$$

where  $\mathbf{n} \in \mathbb{R}^3$ ,  $\|\mathbf{n}\| = 1$ , and  $\mathbf{Q}$  is the projection of  $\mathbf{P}$  onto  $H_{\mathbf{P}}$ ,  $\langle \cdot \rangle$  is the dot product.  $H_{\mathbf{P}}$  is computed by minimizing a local weighted sum of squared distances of the  $\eta$  nearest neighbored map points ( $\mathbf{X}$ ):

$$\arg \min_{\mathbf{Q}, \|\mathbf{n}\|=1} \sum_{i=1}^{\eta} \lambda \langle \mathbf{Q} - \mathbf{X}_i, \mathbf{n} \rangle^2 \quad (4.4)$$

$$\lambda = \theta \|\mathbf{Q} - \mathbf{X}_i\| \quad (4.5)$$

where  $\theta$  is a Gaussian kernel. The weight attached to each neighbored point  $\mathbf{X}_i$  is defined as the function of the distance of  $\mathbf{X}_i$  to the projection of  $\mathbf{P}$  on the plane  $H_{\mathbf{P}}$ , rather than the distance to  $\mathbf{P}$ .

In a second step, a least squared minimization is computed to find a local bi-variate polynomial approximation  $\mathcal{G}$  to the true surface (red surface in Figure 4.10(b)) using  $\eta$  neighborhood point of  $\mathbf{P}$ :

$$\arg \min_{\mathcal{G}} \sum_{i=1}^{\eta} \lambda \|\mathcal{G}(\mathbf{X}'_i) - F_i\|^2 \quad (4.6)$$

where  $\mathbf{X}'_i$  is the projection of the map point  $\mathbf{X}_i$  onto  $H_{\mathbf{P}}$  and  $F_i = \langle \mathbf{X}_i - \mathbf{Q}, \mathbf{n} \rangle$  is the distance of  $\mathbf{X}_i$  to  $H_{\mathbf{P}}$ . This two MLS steps are repeated for every map points to obtain a locally smoothed reconstruction.

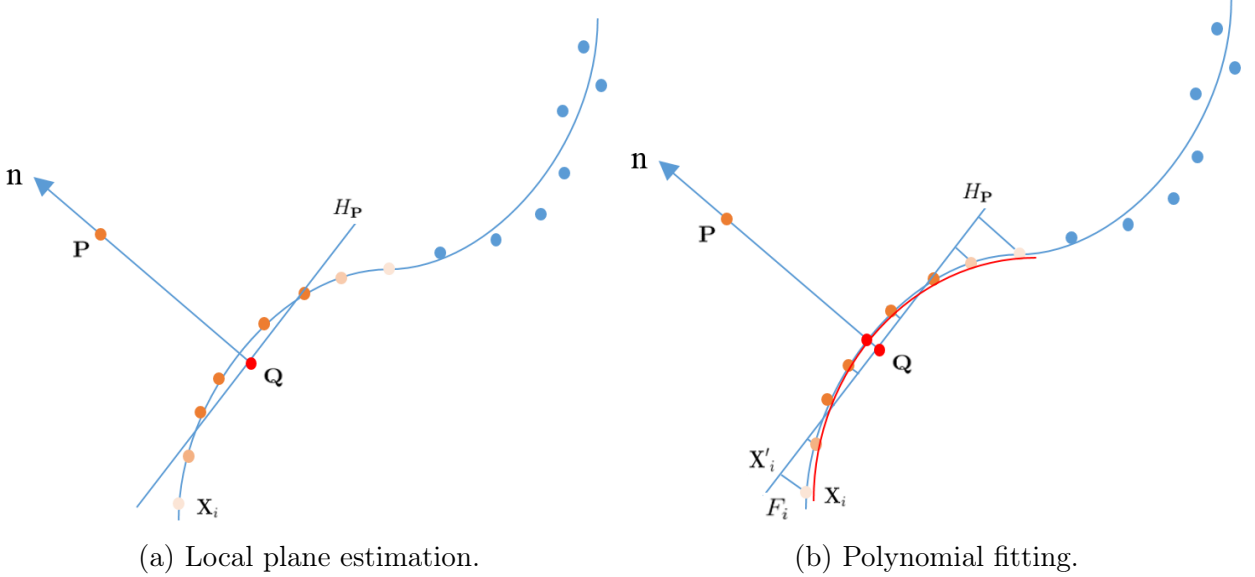


FIGURE 4.10: Surface denoising.

Figure 4.11 shows the reconstruction from different view points after outlier removal and denoising, with normal ORB map point highlighted in red and the newly added points from feature based densification stage are highlighted in blue. Note that there are holes in the reconstruction due to specular reflections and regions of extremely homogeneous texture. To obtain an intensity homogenous reconstruction, it is projected to all keyframes in  $G$ , and we use the average color for every pixel.

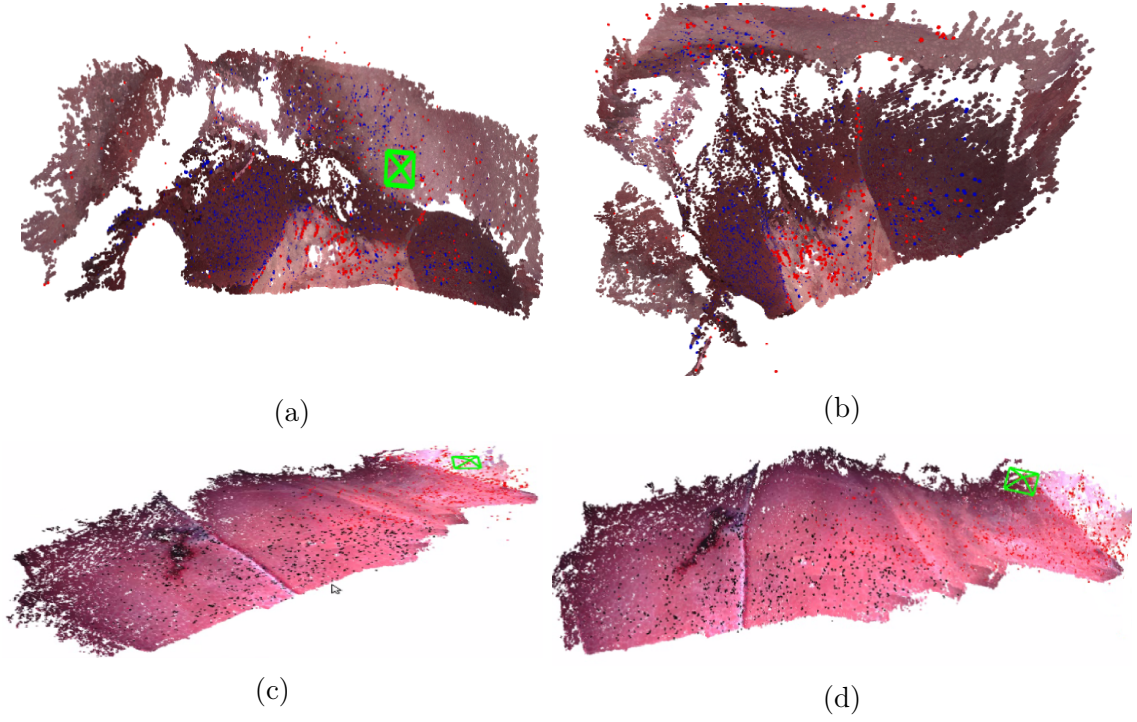


FIGURE 4.11: Denoised reconstructed map from private (a,b) and public (c,d) datasets.

## 4.6 Experimental evaluation

The tracking robustness and reconstruction accuracy of the proposed SLAM system have been quantitatively and qualitatively tested in a series of *in-vivo* experiments. First, we describe in Section 4.6.1 the protocol used to obtain various in-vivo sequences together with CT ground truth. Two pigs were used during this experiment and all guidelines for care and use of animals are followed as indicated in Appendix A. Second, we provide a quantitative evaluation of reconstruction error with respect to CT ground truth in Section 4.6.2. Third, the robustness of system tracking has been qualitatively evaluated with challenging conditions such as organ deformations and partial scene occlusions in Section 4.6.3. Fourth, the parameters settings used during the experiments and processing times are reported in Section 4.6.4. In Section 4.6.5 we present a markerless AR overlay of liver hepatic vein during laparoscope exploration. Additionally, we quantitatively tested system performance on indoor sequences from public dataset. More details of the obtained results can be appreciated in our videos <sup>1</sup>, <sup>2</sup> and <sup>3</sup>.

### 4.6.1 Data acquisition

A live *porcine* experiments were performed inside a CT room (cf. Figure 4.12(a)). In these experiments, a monocular laparoscope was used to explore the abdominal cavity and to record

<sup>1</sup><https://www.youtube.com/watch?v=UzPjHqX5-9A>

<sup>2</sup><https://www.youtube.com/watch?v=R17lsiIRjbM>

<sup>3</sup><https://www.youtube.com/watch?v=oG54CBzqVh0&t=12s>

different exploratory sequences, ranged between 2 to 10 minutes. A CT was then acquired during 10 second expiration breath-hold and while the laparoscope was fixed by means of an articulated arm as shown in Figure 4.12(b). During the CT acquisition, the tip of the laparoscope was included, to be later segmented and extracted from the CT images. The CT images were later segmented manually by an expert to generate a 3D volume with  $0.879\text{mm} \times 0.876\text{mm} \times 0.799\text{mm}$  voxel size and used as our ground truth (cf. Figure 4.12(c)).

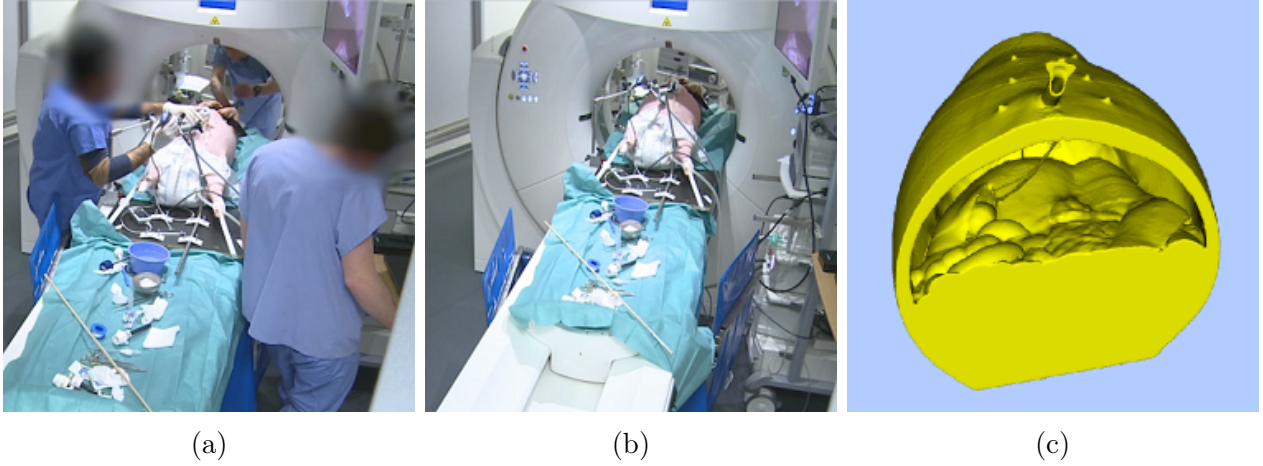


FIGURE 4.12: Data acquisition. (a) Video recording. (b) CT acquisition while laparoscope is fixed and its tip inside the abdominal cavity. (c) Complete CT surface of pig abdominal cavity.

In the recording, we considered two sets of exploratory motions. The first set was to simulate the typical exploration phase of the abdominal cavity that is typically performed by a surgeon at the beginning of the surgery and without any interaction by surgical instruments. The second set is more difficult and challenging, where an instrument is introduced into the abdominal cavity and is used to interact with the liver. Thus, it causes partial scene occlusions and, when interacting with the liver, produces significant organ deformations besides deformation caused by respiration. Additionally, during interaction with liver, we considered two types of deformations that can be classified as: *soft deformation*, caused by simply pushing liver lobe surface and *strong deformation*, caused by changing organ morphology such as lifting the liver lobe.

#### 4.6.2 Quantitative analysis

We use our monocular system to reconstruct a map of liver surroundings abdominal viscera, abdominal wall, and diaphragm from 1min exploratory sequence. During this assessment, we used a CT model as our ground truth and evaluated the reconstruction error of:

1. Tuned ORB-SLAM reconstruction.
2. Reconstruction obtained by densified discrete mapping method obtained by Section 4.4.
3. Quasi dense pairwise reconstruction obtained by Section 4.5.

In order to evaluate the accuracy of SLAM reconstruction with respect to CT surface, we had to recover the real scale and orientation changes, since our monocular reconstruction is up to a similarity transform  $\mathbf{S} \in \mathbf{Sim}(\mathbf{3})$  (i.e. an arbitrary scale and rigid coordinate transform). We compute a best-fitting similarity transform  $\mathbf{S}$  through a Huber robustified non-linear optimization that minimizes the distance between SLAM points to CT surface points:

$$\arg \min_{\mathbf{S} \in \mathbf{Sim}(\mathbf{3})} \sum_j \rho_h(\|\mathbf{Q}_j - \mathbf{S}\mathbf{X}_j\|) \quad (4.7)$$

where  $\mathbf{Q}_j$  is the CT surface point that is closest (i.e. with smallest perpendicular distance) to SLAM point  $\mathbf{X}_j$ . The distance is defined as the Euclidean distance between  $\mathbf{X}_j$  and  $\mathbf{Q}_j$ . This optimization was initialized by manually selecting 3 landmarks to roughly estimate initial scale and orientation by Horn’s algorithm (Horn, 1987). We use Levenberg-Marquardt implemented in *g2o* Kümmerle et al., 2011 to carry out that non-linear optimization. In each iteration, the corresponding points between CT model and SLAM map are recomputed to find the closest CT surface point to each SLAM point (following an ICP scheme). We optimize eq. (4.7) to compute  $\mathbf{S}$ , which aligns SLAM reconstructions to CT model. After convergence, the accuracy was measured by the Euclidean distance between each map point to its closest point on the CT model’s surface and RMSE was used to evaluate the overall error.

#### 4.6.2.1 Sparse reconstruction error

In this section we evaluate the error of the sparse reconstruction of the tuned ORB-SLAM and the improved map obtained by densified discrete mapping method.

1) **Tuned ORB-SLAM reconstruction**, Figure 4.14(a,b) shows the alignment of the reconstructed map (in white) to the CT model (yellow). The overall RMSE was 2.9mm. Figure 4.13 shows distance distributions of the total number of map points, and the accumulative histogram of distances. It can be seen from Figure 4.13(b) that 85% of points with lower distances than 5.3mm. Thus, thresholding errors lower than 5.3mm as inliers (85%) and the rest as outliers (15%), those are the red points in Figure 4.14(c), reduces the RMSE error to 1.9mm. Those 15% outliers were abdominal wall points that are strongly affected by the non-rigid deformation by the breathing cycle and heart beating.

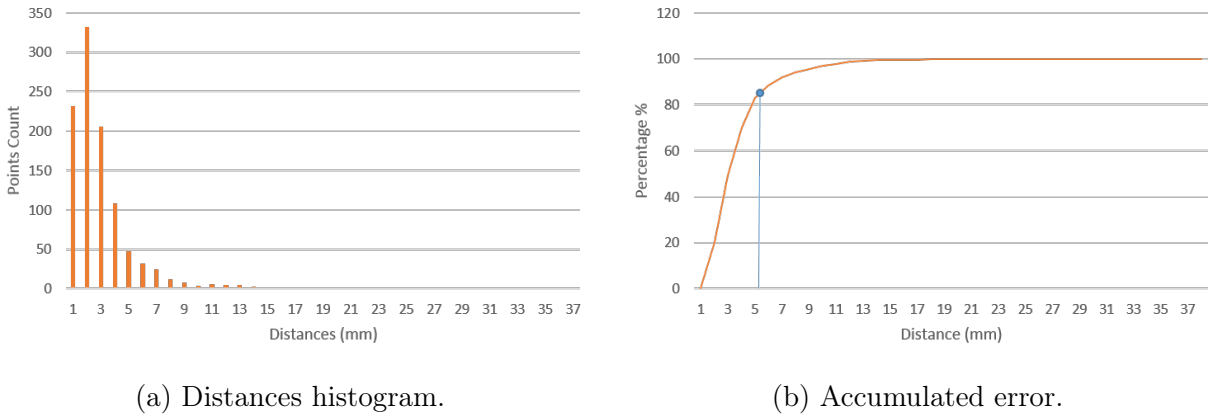


FIGURE 4.13: Distances distributions of tuned ORB-SLAM map.

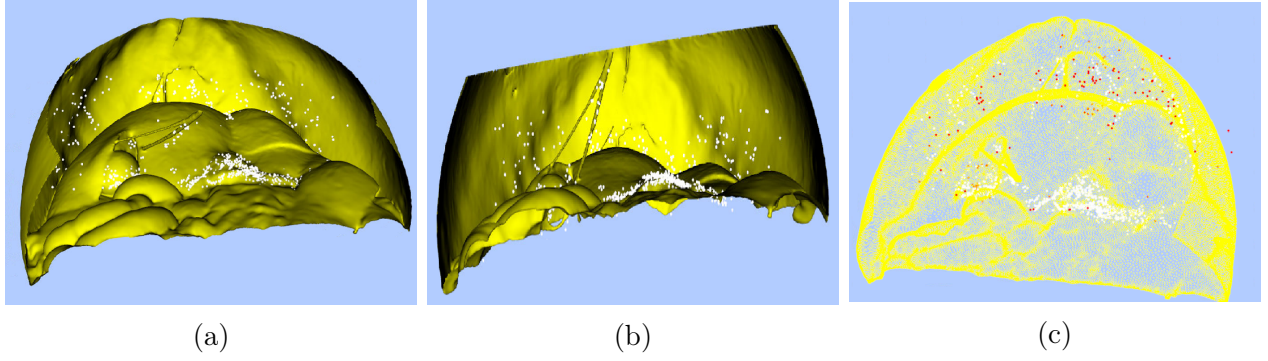


FIGURE 4.14: Alignment of tuned ORB-SLAM map to CT model. (a,b) Alignment from two points of view with the visible part of CT surface in laparoscope images. (d) Outliers rejection, where yellow, white and red points are CT model points, inliers and outliers map points, respectively.

**2) Densified discrete mapping method**, Figure 4.16(a,b) shows the alignment of densified SLAM map to the CT model. The achieved RMSE in this case was 3.5mm. We show in Figure 4.15 the distance distributions of densified map points, in addition to the accumulative histogram of the distances. In Figure 4.15(b) can be seen that 85% of points with lower distances than 5.4mm. Similarly, thresholding errors lower than 5.4mm as inliers (85%) and the rest as outliers (15%), red points indicated in Figure 4.16(c), reduces the overall RMSE error to 2.7mm.

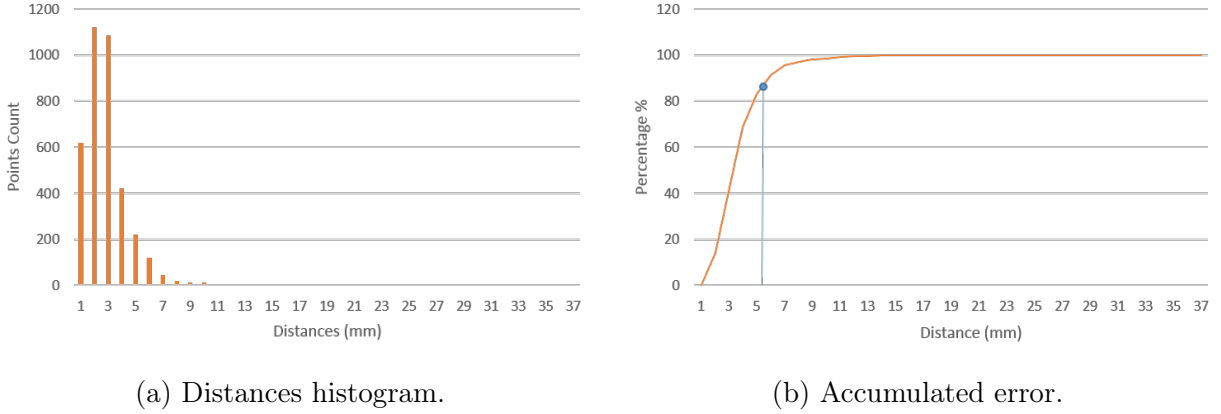


FIGURE 4.15: Distances distributions of discrete densified map.

Therefore, the overall RMSE of final reconstruction of the tuned ORB-SLAM and the densified discrete mapping is 1.9mm and 2.7mm, respectively, excluding points created at deformable areas. As in any SLAM system, camera pose estimation and map are tightly coupled, because map points are involved in camera-only non-linear optimization. Therefore achieving low mapping error is important, to avoid error propagation to camera localization that can lead to drifts.

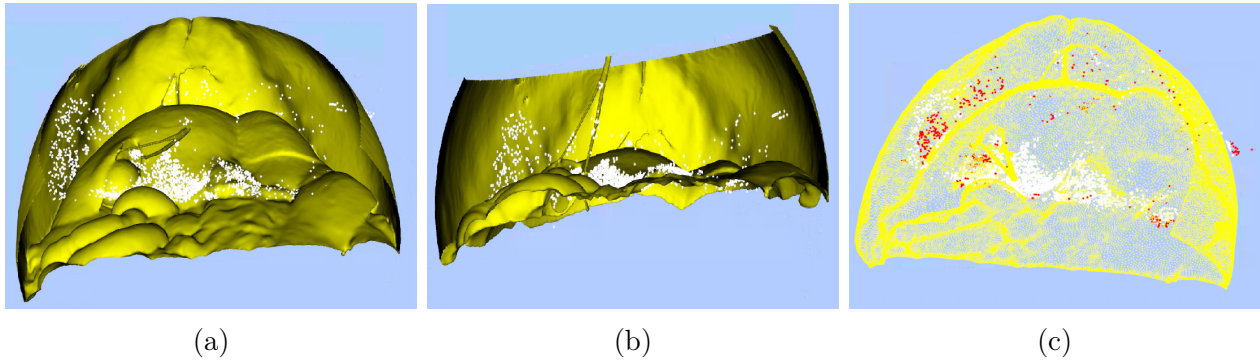


FIGURE 4.16: Alignment of discrete densified map to CT model.

#### 4.6.2.2 Dense reconstruction error

In this section, we evaluated the dense reconstruction approach presented in Section 4.5 with respect to the CT model. The overall RMSE was 4.9mm (cf. Figure 4.18). We also show in Figure 4.17 the distance distributions of dense map, and the accumulative histogram of distance.

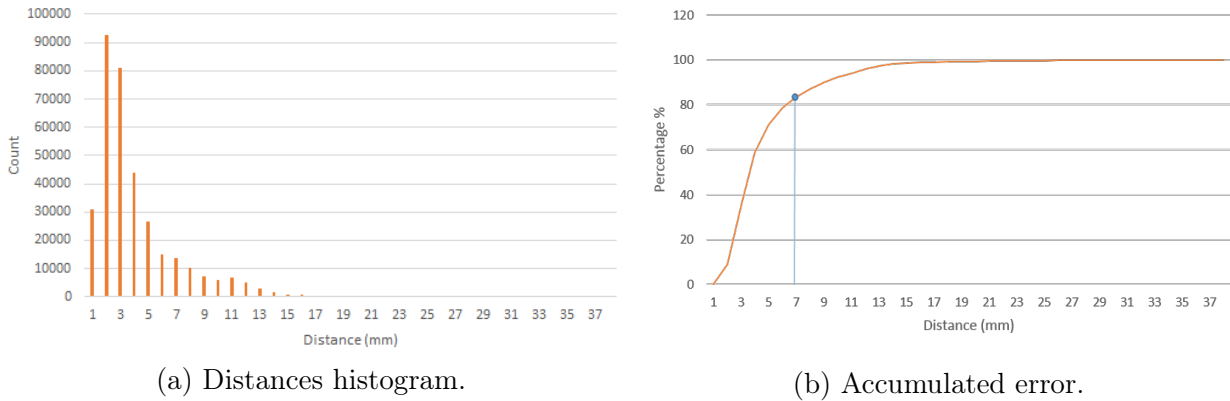


FIGURE 4.17: Distances distributions of dense map.

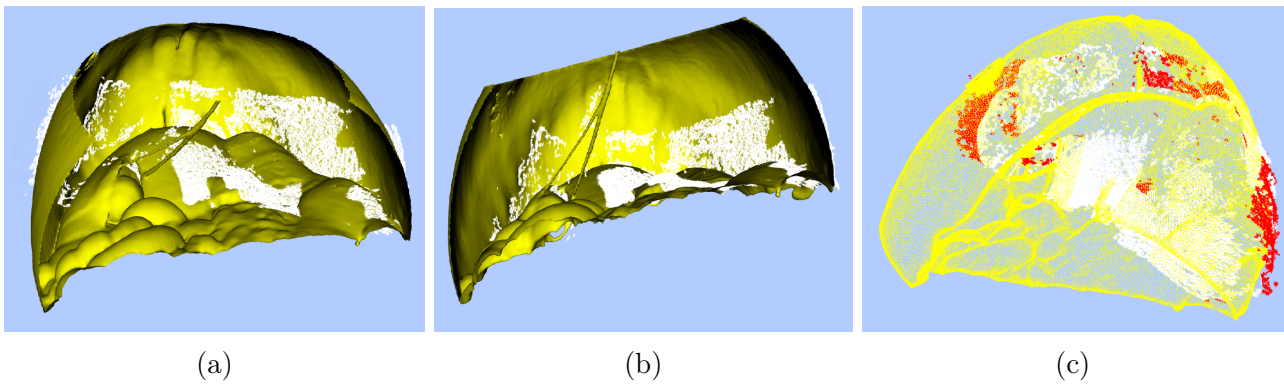


FIGURE 4.18: Alignment of quasi dense map to CT.

In that case 85% of points with lower distances than 6.8mm. Therefore, thresholding errors lower than 6.8mm as inliers (85%) and the rest as outliers (15%), those are red points in Figure 4.18(c), reduces the RMSE error to 2.8mm. The reconstruction density is significantly better than the sparse one, with slightly loss of mapping accuracy where the RMSE of the dense map is 2.8mm and tuned ORB-SLAM map is 1.9mm.

### 4.6.3 Tracking robustness

Due to lack of tracking ground truth, in this section we qualitatively evaluate the robustness of the tracking on different in-vivo sequences with challenging conditions, such as scene occlusions and organ deformations. We recall that the SLAM system uses only ORB points for the camera pose estimation, and we keep this to maintain the real-time tracking and control error propagation in camera-only non-linear optimization. Figure 4.19 shows a robust laparoscope pose estimation with the presence of several occlusions caused by an instrument and/or liver deformations. Figure 4.19(c) shows, from a top view, the estimated laparoscope pose with respect to the reconstructed map. The reconstructed map, keyframes and current laparoscope position for Figure 4.19(d) are displayed in Figure 4.19(e,f). Figure 4.19(g-i) shows the robustness to scene deformation during lifting liver lobe.

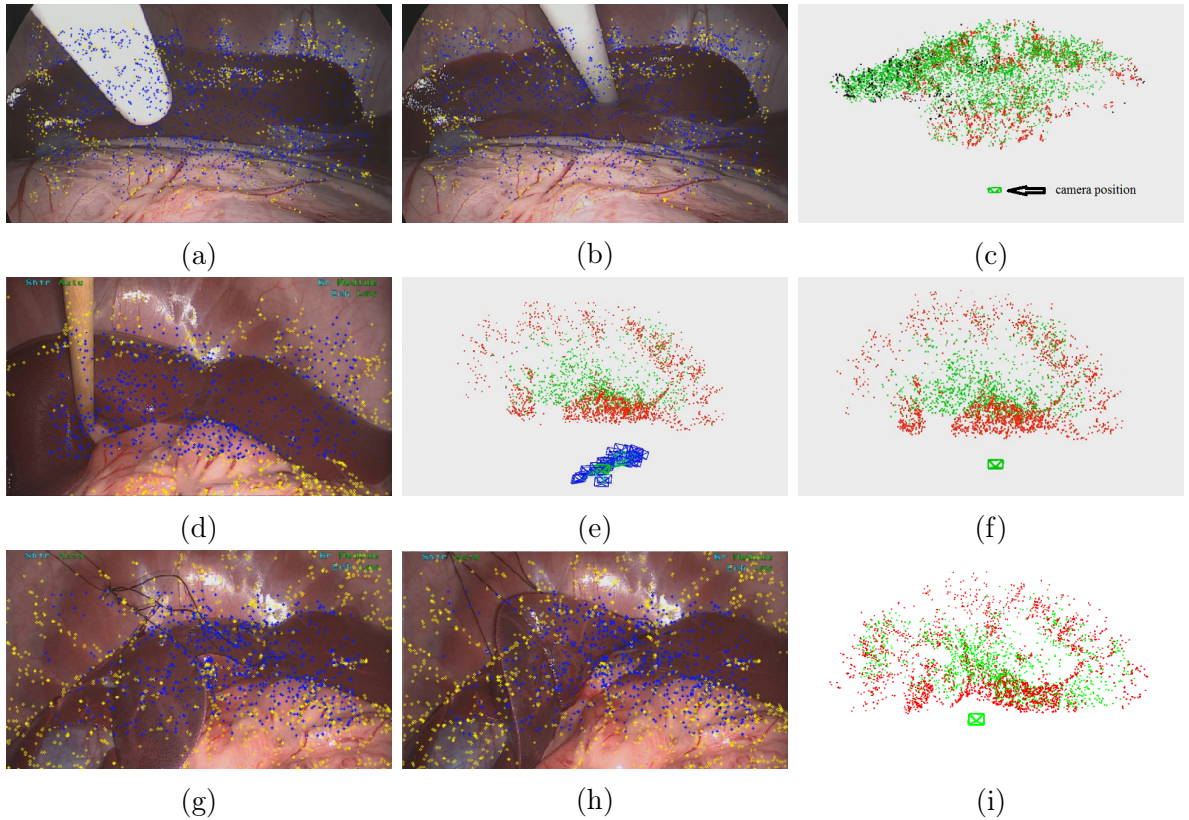


FIGURE 4.19: Endoscope tracking and mapping during partial scene occlusions and deformations.

It is worth noting that the images of the liver of the second pig (cf. Figure 4.19 second and third rows) are extremely difficult due to lack of texture on liver surface, however the

system showed a robust tracking ability, with features only located in the diaphragm wall. The laparoscope pose was successfully tracked during the interaction with the liver in all sequences. In case of tracking failure due to feature deletion during fast laparoscope motions, the system was able to relocate the laparoscope pose once it had moved and few ORB features were detected in the FOV.

#### 4.6.4 Tuning details and computation cost

In this section, we report the computation cost of different system steps used in this chapter. The system has been executed on a desktop PC with Intel Core i7 CPU @2.6 GHz and 4GB RAM. First, in Table 4.1 we report the computation cost of the tuned ORB-SLAM with densified discrete mapping scheme. The system considers ORB points for camera pose estimation, however the tracking time was increased because the optical flow tracking that was allocated in the tracking thread to track features between consecutive frames. Thus, it was necessary to avoid such overhead in the dense pairwise approach, Table 4.2, where the SLAM system is used to collect set of keyframes during exploration phase using only ORB points with average tracking cost 23.4ms. At the end of the exploration phase, the average time required for: feature based densification was 11.2ms per keyframe, depth propagation was 25.1ms per keyframe (time for matching and triangulating points), MLS denoising was 1.1min due to computing normal for each point and polynomial fitting. The total number of points in dense pairwise reconstruction was 348, 068 points (cf. Figure 4.11(a,b)). After obtaining dense scene reconstruction, the laparoscope pose was tracked in 23.4ms on average because only the sparse ORB-SLAM map is used. We provide parameter settings used for dense pairwise reconstruction in Table 4.3.

TABLE 4.1: Average computational cost of ORB-SLAM + dense discrete mapping average

Mapping Thread		Tracking Thread		
ORB mapping	Patch correlation mapping	ORB matching	Lucas-Kanade optical flow	Pose estimation (including optical flow tracking)
158.8ms	5.8ms	6.3ms	33.1ms	60.3ms

TABLE 4.2: Average computational cost of quasi dense pairwise reconstruction.

Pose estimation	Feature based densification	Depth propagation	MLS denoising
23.4ms	11.2ms	25.1ms	1.1min

TABLE 4.3: Tuning parameters used for dense pairwise reconstruction.

$\alpha_1$	$\alpha_2$	$\mathcal{W}$	$\psi$	$\beta$	$\tau$	$\kappa$	$\eta$
0.05	0.09	[15,19]	0.3	1.5	4	40	40

$\alpha$  represents the ratio between neighbored keyframe baseline and scene median depth and we used  $\alpha_1 = 0.05$  and  $\alpha_2 = 0.09$ , which yields a parallax between  $2.9^\circ$  and  $5.1^\circ$  respectively and therefore increases reconstruction accuracy. For ZNCC window, our empirical studies have shown that a good matching result are obtained with a correlation window size  $\mathcal{W}$  between 15 to 19 pixels, where it permits enough textures in the patch and becomes distinguishable. Smaller patch size reduces matching quality where it is highly likely to include only a textureless area. Integral images were used to keep the computation time invariant to the correlation window size as proposed by Stoyanov et al., 2010. We used a low correlation score  $\psi$  to enable more matches and thus reduces the false negative matches, whereas the filtering and denoising step can handle the false positive ones.  $\beta$  was set to 1.5 pixels and used to determine the smoothness of the disparity map and is determined adaptively depending on the color similarity and proximity between the seed and candidate pixels. For removing outliers,  $\kappa$  neighbored points are used by the statistical filter and we set  $\kappa = 40$  points, with a threshold  $\tau = 4$  multiplied by standard deviation of all points distances. Large values of  $\kappa$  and  $\tau$  are used to ensure points very far from true surface are eliminated and thus bias in MLS surface estimation. We use the same number of neighbored points to each point in MLS polynomial fitting  $\eta = 40$ .

#### 4.6.5 AR superimposition of intra-operative CT models

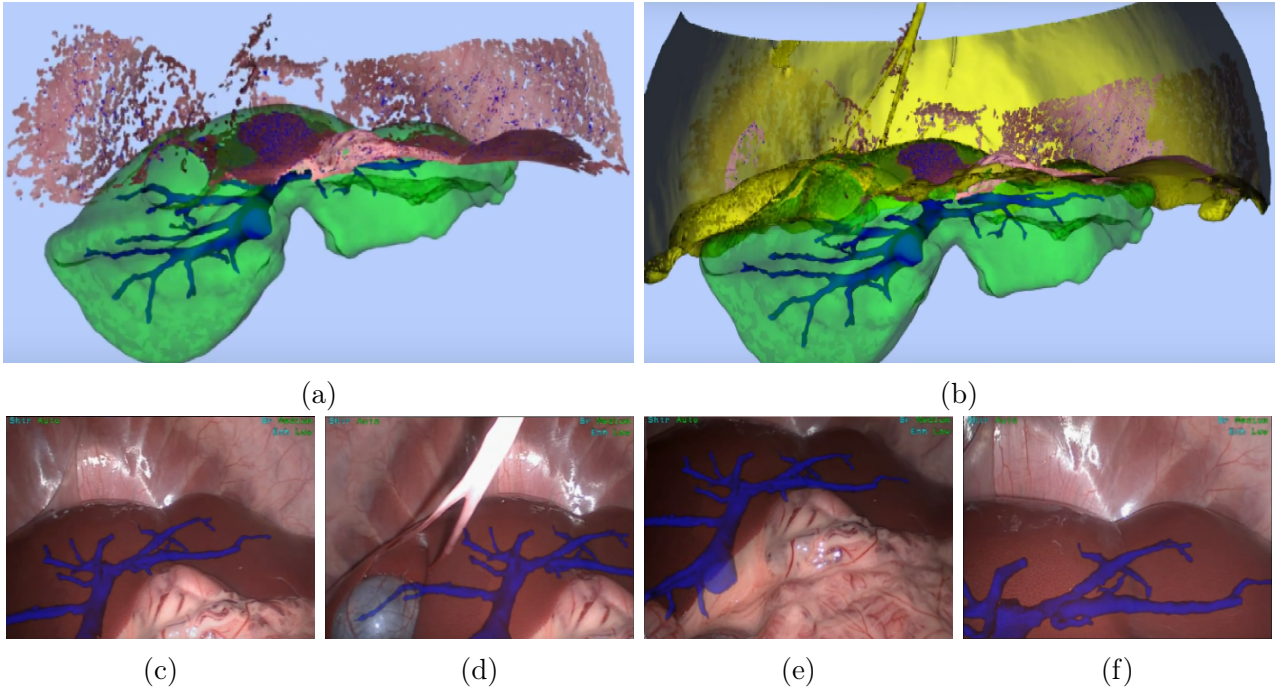


FIGURE 4.20: Markerless AR overlay of liver hepatic vein. (a,b) Alignment of hepatic vein (blue), liver (green) and abdominal wall (yellow). (c-f) Image samples of AR rendering during exploration

The rich (i.e. including original scene texture) and dense scene representation facilitates the registration of the intra-operative CT models. These models are obtained by intra-operative CT acquisition then manually/automatically segmented. For AR overlay of these models onto

endoscopic images, the only needed is the similarity transform that align them with the dense SLAM map. Figure 4.20(a,b) shows the alignment between liver hepatic vein (blue), liver surface (in green), abdominal wall(in yellow) and the dense map using the similarity transform obtained from Section 4.6.2.2. After registration is finished, the relative laparoscope camera pose is tracked by the SLAM system in real time, where only ORB points are considered for tracking. We follow the same AR visualization pipeline used in Section 3.4.3, where a virtual camera is located at the estimated camera pose by the SLAM system and capture virtual image of intra-operative model. This image is later transparently fused with real laparoscope image. This AR visualization is performed in real-time and we show in Figure 4.20(c-f) frame samples of the markerless AR overlay of the hepatic vein in a transparent representation during laparoscope exploration.

#### 4.6.6 Performance on indoor sequences

The TUM RGB-D dataset (Sturm et al., 2012) contains indoors sequences from RGB-D sensors grouped in several categories to evaluate object reconstruction and SLAM/odometry methods under different texture, illumination and structure conditions. We show the pairwise dense reconstruction results on a subset of the sequences that most RGB-D methods usually use. Figure 4.21(a,b,d) shows image samples of the processed sequence together with the obtained dense reconstruction results. The sequences have been recorded from one or two meters distance camera and ranged in difficulty from orthogonal zig-zag (cf. Figure 4.21(d)) to highly curvature (cf. Figure 4.21(a-c)) structures. The blue frustums shows the trajectory of the hand-held camera in each sequence. Furthermore, the system has been tested on a very challenging indoor sequence of a person sitting while a hand-held camera moves around his face Figure 4.21(c). This sequence has been recorded inside a room with different light sources, i.e. fixed light source and extra lights comes from a room window. Despite human faces are challenging to reconstruct, due to lack of features on skin, the system was able to provide a promising face reconstruction. It is worth noting that we have processed each of these sequence with the same tuning defined in Table 4.3. As can be seen from Figure 4.21, the density of the reconstructed map is significantly improved using the proposed pairwise dense reconstruction approach.

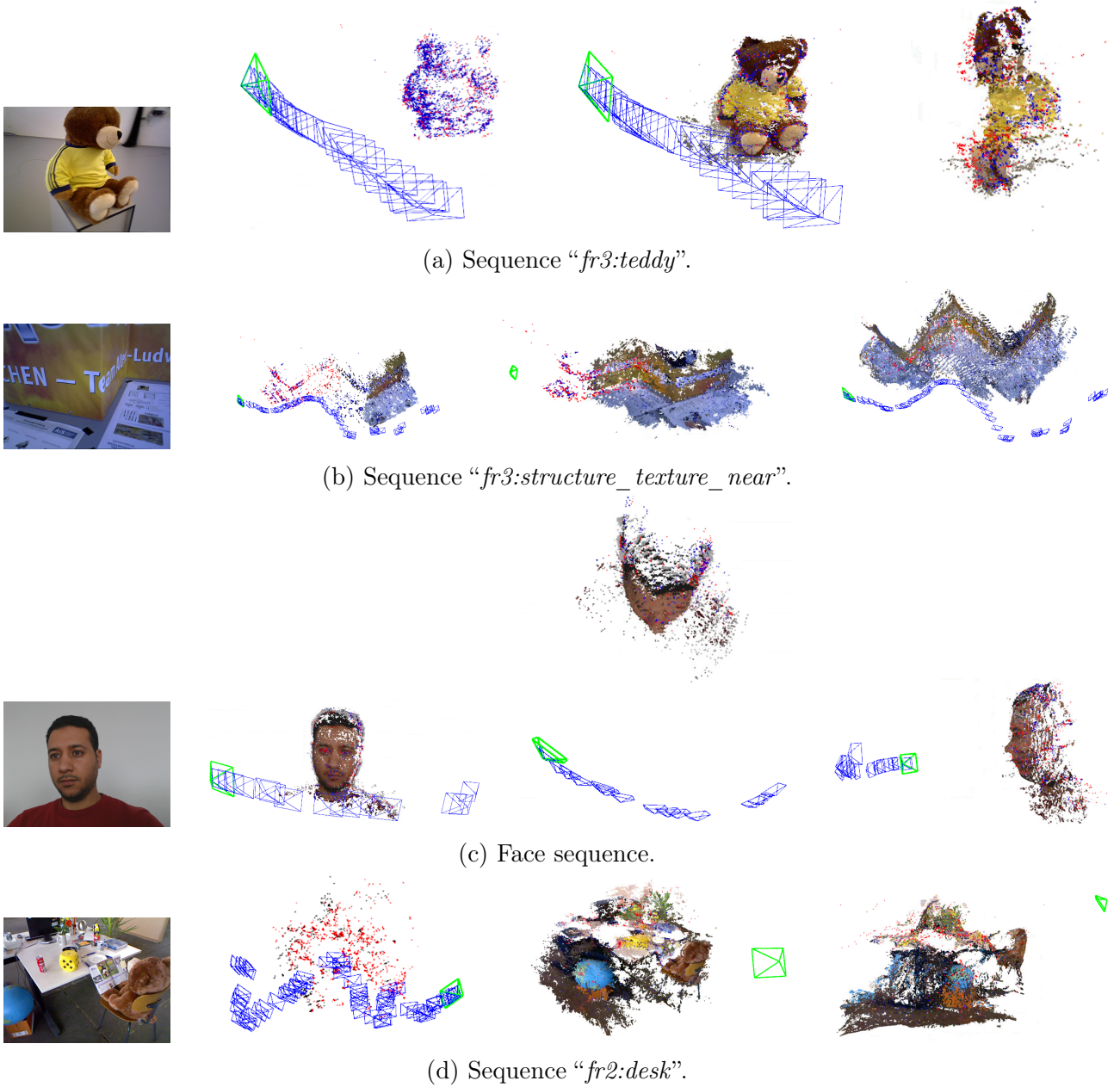


FIGURE 4.21: Pairwise dense reconstruction of different indoor sequences from public dataset (Sturm et al., 2012).

Table 4.4 reports a quantitative evaluation with respect to ground truth information available at TUM dataset Sturm et al., 2012, except for Face sequence because no ground truth was available. This ground truth consists of depth maps for each frame in the sequences, which is obtained by RGB-D sensor. For each processed sequence, our system selects a pair of keyframes for stereo matching, and estimate the depth for set of corresponding pixels. We considered depth map of one of these two keyframes as our ground truth, and computed the RMSE of the Euclidean distances for all pixels that have depth estimation by our system and RGB-D sensor. We provide in Table 5.5 the averaged RMSE, in centimeter, for all processed keyframes

in the sequences. Before the comparison, we had to estimate the monocular scale factor, and we have done that by means of Least Median of Squares. Sequence “*fr2:desk*” contains vast and very homogenous textureless regions such as desk surface, floor, and background, thus the associated error is high.

Sequence	Average RMSE (cm)
<i>fr3:teddy</i>	4.5
<i>fr3:structure_texture_near</i>	7.8
<i>fr2:desk</i>	23.3

TABLE 4.4: Average reconstruction error with respect to RGB-D sensor.

## 4.7 Conclusion

In this chapter, in an initial step we have proved that with careful re-tuning of different ORB-SLAM parameters, a robust endoscope camera tracking can be achieved even with the presence of scene occlusion and hard organ deformations. The fact that the system relies on salient image feature for tracking and mapping tasks enables real-time performance and gain robustness to challenges exist in endoscopy such as illumination and orientation changes, in addition to repetitive textures. The system performs the tracking and sparse scene reconstruction tasks using the only input of image stream gathered by a monocular endoscope. It also shows a remarkable re-localization performance to recover the endoscope pose after tracking failure and when extracting/re-inserting the endoscope to the abdominal cavity.

In a second step, we have extended the SLAM system with a guided patch correlation search to tackle the key factor limiting its mapping performance, and hence yielding a better discrete map with increased number of feature points. However, it concentrate the reconstruction at feature image locations, which are not sufficient to describe surgical scene.

In a third step, we have presented a simple yet but effective approach for quasi-dense pairwise reconstruction of MIS scenes, that significantly and accurately densifies the sparse SLAM reconstruction. The proposed approach processes pairs of registered SLAM keyframes with dense stereo method. Although densification is embedded in SLAM pipeline, we keep all the benefits of state-of-the-art SLAM system, including fast tracking, mapping and automatic relocalization, where only sparse features are used for all SLAM tasks. Our results on in-vivo porcine dataset were very promising, with a RMSE of 2.8mm, when excluding the outliers in deformable areas. The dense scene reconstruction has been an important element for registering intra-operative CT models. Benefiting from the real time estimation of the relative endoscope pose with respect to the dense reconstruction, we showed a markerless AR overlay of the liver hidden structure. Furthermore, the proposed pairwise dense reconstruction system has been quantitatively tested on different indoor sequences from public dataset and showed a significant improvement in the reconstructed map with good accuracy.

Despite the nice benefits of the obtained dense SLAM system, it still has major drawbacks that includes: 1) Long waiting time till scene exploration is finished, and acquiring enough keyframes before densification begin; 3) The denoising step is blindly smoothing the reconstructed surface to fit a local plane to each point using set of nearest neighbors points, thus fits planar and less curved surfaces, whereas high errors can arise at surface discontinuities; 3)

This denoising step takes nearly 1-2min depending on the density of the reconstruction. Consequently, in next chapter we present a novel live dense SLAM system that is able to operate the tracking in real-time while incrementally computing the dense reconstruction of the surgical scene.



# Chapter 5

## Live Tracking and Dense Reconstruction For MIS

### Contents

---

<b>5.1</b>	<b>Introduction</b>	<b>70</b>
<b>5.2</b>	<b>Approach overview</b>	<b>71</b>
<b>5.3</b>	<b>Frames cluster selection for dense reconstruction and cluster bundle adjustment</b>	<b>72</b>
<b>5.4</b>	<b>Reconstruction of a keyframe's depth map</b>	<b>73</b>
5.4.1	The variational formulation	73
5.4.2	ZNCC data term	73
5.4.3	The regularizer	74
5.4.4	Initialization	75
5.4.5	Energy minimization	76
<b>5.5</b>	<b>Live alignment of keyframe depth maps</b>	<b>78</b>
<b>5.6</b>	<b>Experimental Results</b>	<b>78</b>
5.6.1	Benchmark hardware and compared methods	78
5.6.2	Datasets	78
5.6.3	Quantitative evaluation using dense stereo	79
5.6.4	Qualitative evaluation on patient data	87
5.6.5	Free parameters tuning	88
5.6.6	Processing time	88
5.6.7	Augmented reality annotations	89
5.6.8	Performance on indoor sequences	90
<b>5.7</b>	<b>Conclusion</b>	<b>92</b>

---

## 5.1 Introduction

Recently, SLAM approaches have received a significant boost in performance to cross the border of only localizing the camera, and to provide a visually appealing 3D representation of the observed scene. Dense or semi-dense SLAM approaches (Newcombe et al., 2011a; Engel et al., 2014; Pizzoli et al., 2014; Concha et al., 2015; Engel et al., 2018) achieves high quality dense reconstruction results in real-time. Newcombe et al., 2010 showed the advantage of reconstruction from large number of video frames taken from very close viewpoints, where photometric-consistency is possible. Newcombe et al., 2010 enforced a smoothness priors over the reconstructed scene by minimizing a regularized energy functional based on aggregating a photometric cost over different depth hypothesis and penalizing non-smooth surfaces. The obtained dense model has improved the tracking robustness in case of motion blur or low textured scenes (Newcombe et al., 2011a). Pizzoli et al., 2014 have introduced a probabilistic approach for dense reconstruction by combining Bayesian estimation and convex optimization of Newcombe et al., 2010, to take into account the uncertainty in measurements, enforce spatial regularity, and to mitigate the effect of noisy camera localization. However, the accuracy of reconstruction degrades in texture-less areas. Thus, Concha et al., 2015 has incorporated superpixels (i.e. planar areas) and indoor scene understanding in the dense reconstruction to tackle the problem of textureless areas.

Despite the ground-breaking results of these approaches, they have been of limited use in endoscopy, where they require a constant illumination and unchanged pixel brightness with respect to the view direction. Despite these assumptions, they have been experimentally proven to perform robustly for indoor scenes. However, these assumptions are violated in endoscopy where the light source is attached to the endoscope tip, which produces significant illumination changes as the endoscope explores the scenes, in addition to specular reflection. Moreover, these approaches are equally weighting the measurements of pixel depths from small and large parallaxes, distant and close scene points. Furthermore, an inadequate number of images can lead to a poorly constrained initialization for the optimization and erroneous measurements. The lack of a unified criterion for image selection depending on the motion of the camera and scene structure.

In contrast to current SLAM dense approaches, in this chapter we present a novel real-time dense SLAM system that is able to cope with challenges in endoscopy and has been successfully applied in laparoscopy. The proposed system extends the work presented in previous chapter, with a novel dense multi-view stereo-like approach for recovering dense scene geometry. The system is extended in several important ways. Firstly, a new thread is added to the system performing the dense scene reconstruction that runs live and in parallel with ORB-SLAM tracking and mapping threads without interrupting them, to maintain real-time tracking. This eliminates the wait for the abdominal cavity exploration to finish before densification. Secondly, only important keyframes images are selected for densification, and around each keyframe a cluster of neighbored frames are selected according to parallax criteria and their relative poses are accurately computed for a high quality depth estimation of every single pixel in those keyframes. Due to the effective selection of video frames based on parallax criteria, the proposed method can outperform the pure stereo based reconstruction, because the frames cluster can provide larger parallax from the endoscope’s motion.

The crux of the dense reconstruction is a variational approach, inspired by Newcombe et

al., 2011a, with a Huber norm regularizer and illumination invariant data term. To design a robust data term with illumination variability in endoscopy, one usually has two options: the first option is to model the light source as proposed by Collins et al., 2012a and Engel et al., 2018. Our approach follows a second option by considering an illumination invariant image representation (Chang et al., 2014; Marcinczak et al., 2014). Marcinczak et al., 2014 transform image to illumination invariant representation, however the data term relies on measuring pixels similarities, that is very sensitive to minor transformation, both in geometry (shifts and rotation) and in imaging conditions (noise and blurring). Consequently, Chang et al., 2014 considered the use of ZNCC to gain more tolerance to different camera gain or bias and provide better fidelity in textureless regions with stereo scope. Both Chang et al., 2014; Marcinczak et al., 2014 provides only local reconstruction of the visible region in either a stereo pair or a reference monocular image, but not a global and complete reconstruction of all captured regions in the surgical scene. For global reconstruction Turan et al., 2017 proposed to fuse several depth maps obtained by SfS, however it is only validated on synthetic dataset. In this chapter, we propose a dense SLAM system that provides a live global and consistent dense reconstruction of the surgical scene by merging and aligning keyframe’s depth maps on-line. The work presented in this chapter has been submitted to IEEE Transaction on Medical Imaging Journal (Mahmoud et al., 2018).

## 5.2 Approach overview

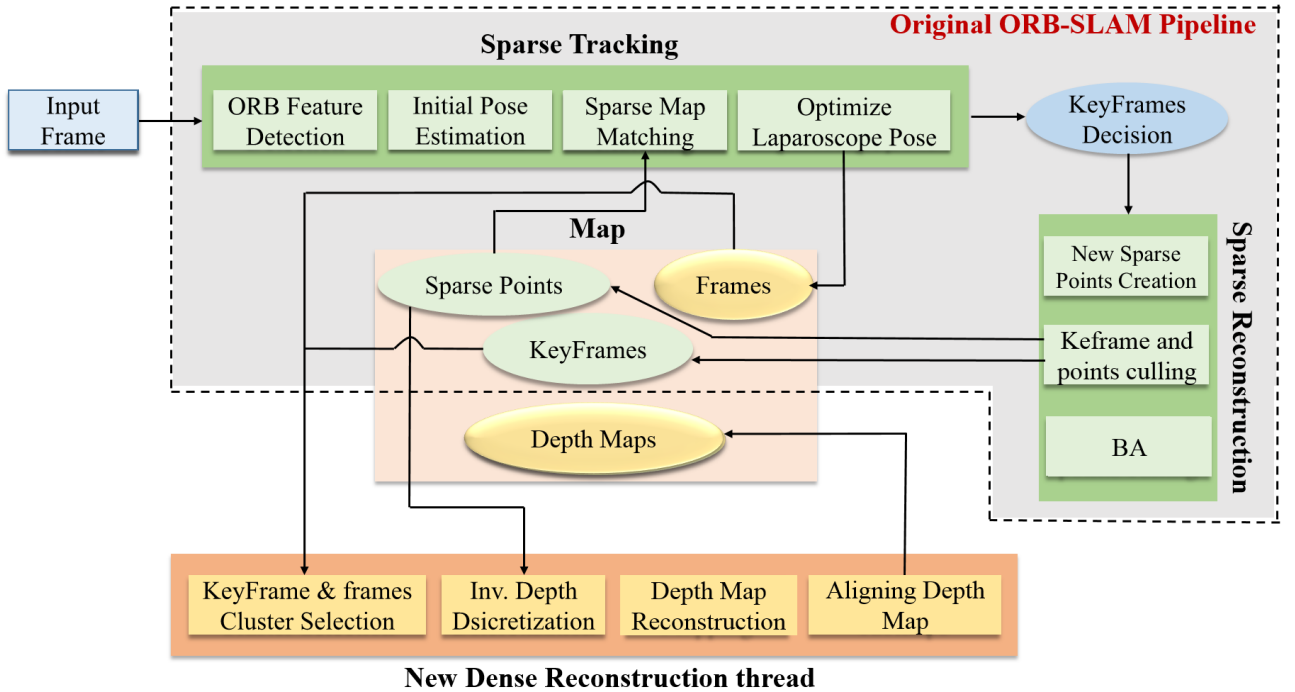


FIGURE 5.1: System Architecture.

We outline our approach in Figure 5.1. We note that our system is applicable to any movable endoscope with a monocular camera, but here we focus on monocular laparoscopes. We assume

laparoscope is pre-calibrated with fixed intrinsic and lens distortion has been compensated. We give the default values for all free parameters in Section 5.6.5. The sparse ORB-SLAM is extended with a new thread for dense reconstruction. The dense reconstruction thread consists of four sequential modules. In the first module, we select a subset of keyframes, during live camera tracking, among all available keyframes. For each considered keyframe ( $I_r$ ), a cluster of neighbor frames  $\{I_1 \dots I_n\}$  is selected to have partially overlapping surface visibility (cf. Section 5.3). In the second module, we exploit the sparse reconstruction to define the range of depths used to construct a 3D cost volume (cf. Section 5.4.4). In the third module, we perform dense reconstruction for each selected keyframe using a variational approach based on Newcombe et al., 2011a. We differ by minimizing a global energy with an illumination-invariant ZNCC data term and Huber norm regularizer (cf. Section 5.4.5). In the fourth module, we obtain a globally consistent reconstruction by aligning the keyframe depth maps with the sparse SLAM map (cf. Section 5.5). The scene is incrementally densified on-line and without interrupting the live endoscope tracking.

### 5.3 Frames cluster selection for dense reconstruction and cluster bundle adjustment

Our dense reconstruction thread aims at estimating the depth map (i.e. depth of every pixel) of a subset of selected keyframes. This can be computationally expensive, so we automatically choose only a subset of keyframes to densify. The selection criterion is the visibility of the current dense reconstruction in a given keyframe  $I_r$ . This visibility is determined by projecting the current dense reconstruction to  $I_r$ , and if the visible fraction is below 50%,  $I_r$  is selected for densification.

Upon selecting  $I_r$ , we define a cluster of  $n$  neighbor frames,  $\{I_{i_1} \dots I_{i_n}\}$ . The criterion for including the frames in the cluster is a measure of parallax. This is defined as the ratio between the sparse SLAM points median depth and the baseline between  $I_r$  and  $I_{i_n}$ . Frames are stored according to their temporal location in the sequence. We then search for the most extreme frame to  $I_r$  whose parallax exceeds a threshold  $\alpha_1$ . This extreme frame and all intermediate frames are added to the cluster. The threshold  $\alpha_1$  controls the tradeoff between depth accuracy and frames overlap, where a small  $\alpha_1$  leads to noisy depth maps, but a higher value reduces the percentage of the overlapping pixels. It also balances the rendered parallax with photometric distortion caused by strong viewpoint change, and computation time as more frames needs to be processed. In a second stage, frames in the cluster are reduced by removing frames from the cluster with low relative parallax because they are not informative, thus it is important keep only the most informative ones and hence avoid longer computation times. The condition applied is that if the parallax between frame  $I_{i_m}$  and its neighbors  $I_{i_{m-1}}$  and  $I_{i_{m+1}}$  is lower than a  $\alpha_2$  threshold, frame  $I_{i_m}$  is removed from the cluster.

ORB-SLAM estimates frame poses when they are grabbed, but those poses are not updated afterwards, unlike the keyframe poses estimation which are continuously refined in the BA. Hence, the estimated poses of the frames in the cluster are not accurate because they are not included in the BA of ORB-SLAM. We re-estimate those poses accurately by a full BA that uses the tracked features from ORB-SLAM and minimizes eq. 5.1 across all the frames in the cluster and some of the other ORB-SLAM keyframes (up to 15 keyframe). The keyframes are

selected as those with the most features common to  $I_r$ .

$$\underset{\mathbf{T}_i, \mathbf{X}_j}{\operatorname{argmin}} \sum_{i,j} \rho h(\|\mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{X}_j)\|^2) \quad (5.1)$$

The index  $i$  ranges over all images in the frames cluster and selected SLAM keyframes, and  $j$  ranges over feature points observed by more than two cameras in the BA. The global reference is fixed during the BA to the keyframe  $I_r$ . We use Levenberg-Marquardt implemented in *g2o* Kümmerle et al., 2011 to carry out that BA. The result of this computation is a set of relative poses  $\{\mathbf{T}_{i_1,r} \dots \mathbf{T}_{i_n,r}\}$  from  $I_r$  to  $\{I_{i_1} \dots I_{i_n}\}$ .

## 5.4 Reconstruction of a keyframe’s depth map

### 5.4.1 The variational formulation

We proposed a variational energy minimization to estimate the inverse depth map  $\rho(\mathbf{x}) : \Omega \rightarrow \mathbb{R}$  for a given keyframe image  $I_r$ . We use grayscale image, denoted by  $I_r : \Omega \rightarrow \mathbb{R}$ , where  $\Omega \subset \mathbb{R}^2$  is the 2D image domain. Our energy is the sum of a regularization term  $R(\mathbf{x}, \rho(\mathbf{x}))$ , and a weighted ZNCC data term  $C(\mathbf{x}, \rho(\mathbf{x}))$  with the following form:

$$\begin{aligned} E(\rho) &= \int_{\Omega} \{\lambda(\mathbf{x})C(\mathbf{x}, \rho(\mathbf{x})) + R(\mathbf{x}, \rho(\mathbf{x}))\} d\mathbf{x} \\ \lambda(\mathbf{x}) &\triangleq \lambda \rho(\mathbf{x}) \end{aligned} \quad (5.2)$$

where  $\lambda$  is a constant and  $\lambda(\mathbf{x})$  is a spatially-varying weighting factor that determines importance of the data term of pixel  $\mathbf{x}$ . Our empirical studies have shown that the geometrical accuracy of the recovered depth is lower for distant scene points than for closer ones because they generally have lower parallax. Thus, differently from Newcombe et al., 2011a, we scale the weight by  $\rho(\mathbf{x})$  to reduce the data term strength for distant points.

To avoid introducing outliers in the dense reconstruction, we first detect specular reflections in  $I_r$ . This is done by thresholding saturation in HSV space with a free parameter  $\tau$ , similar to Section 4.5.2. All pixels in these areas are eliminated before the optimization, because there is high uncertainty in their estimated depths.

### 5.4.2 ZNCC data term

In Newcombe et al., 2011a; Concha et al., 2015; Pizzoli et al., 2014 a per-pixel Sum of Absolute Difference (SAD) of intensity values across a cluster of images is used. In contrast, our data term is based on the ZNCC over a window around each pixel, summed for all the images in the cluster, to obtain an illumination invariant data term that can cope with the severe illumination variability in endoscopy and to achieve tolerance to endoscope gain or bias. Each pixel  $\mathbf{x} = (u, v)^T \in \Omega$  in  $I_r$  is first back-projected using  $\rho(\mathbf{x})$  in the coordinate system of  $I_r$ :

$$\mathbf{X} = h^{-1}(\mathbf{x}, \rho(\mathbf{x})) \quad (5.3)$$

$$h^{-1}(\mathbf{x}, \rho(\mathbf{x})) \triangleq \frac{1}{\rho(\mathbf{x})} \mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (5.4)$$

We then project  $\mathbf{X}$  to each frame  $I_i$  in the cluster  $\{I_{i_1} \dots I_{i_n}\}$ , denoted by the 2D point  $\mathbf{x}_i$ :

$$\mathbf{x}_i = \pi(\mathbf{T}_{i,r}, \mathbf{X}) \quad (5.5)$$

where  $\mathbf{T}_{i,r}$  is the transformation from the reference keyframe  $I_r$  to frame  $I_i$ , computed by BA in Section 5.3, and  $\pi$  is the same as eq. (3.6). The data term  $C(\mathbf{x}, \rho(\mathbf{x}))$  is computed by projecting pixel  $\mathbf{x}$  in the reference image  $I_r$  onto  $I_i \in \{I_{i_1} \dots I_{i_n}\}$  using eq. (5.5), and a ZNCC with correlation window of size  $\mathcal{W}$ :

$$C(\mathbf{x}, \rho(\mathbf{x})) = \frac{-1}{n} \sum_{l=1}^n \text{ZNCC}(I_r(\mathbf{x}), I_{i_l}(\mathbf{x}_{i_l})) \quad (5.6)$$

The pixels that are non-visible in all cluster frames (i.e. projected outside the image dimension) are assigned zero in the data term and eliminated before the optimization to avoid inaccurate estimation of their depths. Those ignored pixels are highly likely to be reconstructed from another reference keyframe if they become visible. Additionally, we threshold the ZNCC by  $\psi$  to detect occlusions and to improve the accuracy of the reconstruction at depth discontinuities, unlike Newcombe et al., 2011a that used  $L_1$  norm.

### 5.4.3 The regularizer

We use a regularizer term  $R(\mathbf{x}, \rho(\mathbf{x}))$ . To enable a smoother reconstruction of the scene, but also to preserve depth discontinuities. This is achieved with a weighted Huber norm over the gradient of the inverse depth image:

$$R(\mathbf{x}, \rho(\mathbf{x})) = g(\mathbf{x}) \|\nabla \rho(\mathbf{x})\|_{\in} \quad (5.7)$$

$$\|\cdot\|_{\in} = \begin{cases} \frac{\|\cdot\|_2^2}{2\in}, & \text{if } \|\cdot\|_2 \leq \in \\ \|\cdot\|_1 - \frac{\in}{2}, & \text{otherwise.} \end{cases}$$

where  $\in$  is a free parameter of the Huber norm which determines when  $L^1$  forming Total Variation (**TV**) or  $L^2$  norm are used (Newcombe et al., 2011a), to reduce the effect of the undesired stair-casing resultant from a pure **TV**. To maintain depth discontinuities across image edges, we use a per-pixel weight  $g(\mathbf{x})$  that decreases the regularization strength at high gradient pixels in the reference keyframe  $I_r$ :

$$g(\mathbf{x}) = e^{-\omega \|\nabla I_r(\mathbf{x})\|_2} \quad (5.8)$$

where  $\omega$  is a free parameter.

#### 5.4.4 Initialization

The ZNCC data term  $C(\mathbf{x}, \rho(\mathbf{x}))$  is evaluated for keyframe  $I_r$  by means of a 3D cost volume as shown in Figure 5.2. This has dimension  $M \times N \times \xi$ , where  $M \times N$  is the image resolution of  $I_r$  and  $\xi$  is number of points sampling the inverse depth, that ranges between  $\rho_{min}$  and  $\rho_{max}$ . This cost volume is computed only once and an initial depth map is estimated from the cost volume by selecting  $\rho(\mathbf{x})$  that minimize eq. (5.6) for each pixel  $\mathbf{x}$ . This is performed with an exhaustive search optimization over the range of inverse depths  $[\rho_{min}, \rho_{max}]$ .

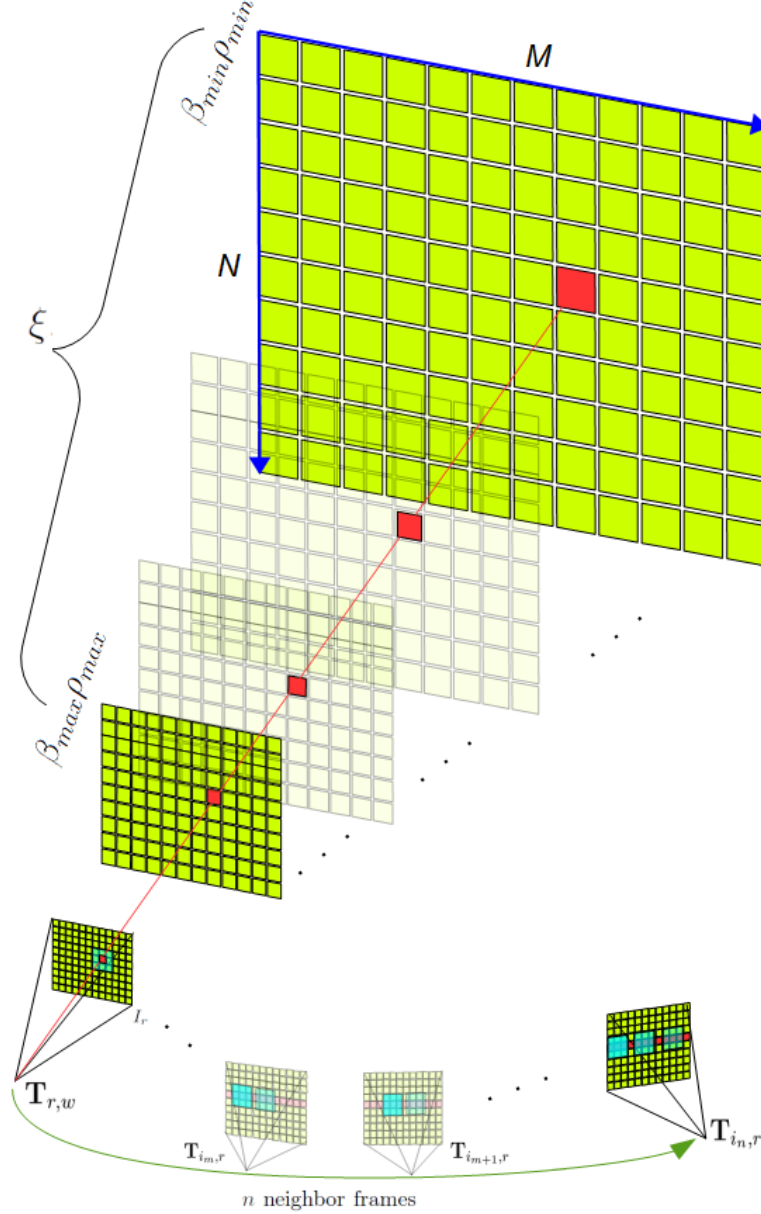


FIGURE 5.2: Cost volume construction of reference keyframe  $I_r$  with relative pose  $\mathbf{T}_{r,w}$  with respect to SLAM map  $w$  and  $n$  neighbor frames with relative pose  $\mathbf{T}_{i_n,r}$  with respect to  $I_r$ . Each pixel in  $I_r$  has an associated row of entries in cost volume (shown in red) that store the average ZNCC cost computed for the corresponding  $\rho \in [\beta_{min}\rho_{min}, \beta_{max}\rho_{max}]$ .

To define  $[\rho_{min}, \rho_{max}]$ , we exploit the scene depths provided by the sparse SLAM map, and compute a histogram of inverse depths of all visible sparse map points in  $I_r$  (i.e. projected inside  $I_r$ ). To be robust to outliers exist in the sparse reconstruction, the 20% extreme closer and farther depths are ignored. Additionally, to include the extreme points which may have been incorrectly excluded, this interval is extended with two empirical factors  $\beta_{min}$  and  $\beta_{max}$  yielding the final interval as  $[\beta_{min}\rho_{min}, \beta_{max}\rho_{max}]$ . This range of inverse depths is evenly discretized into  $\xi$  sampling points.

### 5.4.5 Energy minimization

eq. (5.2) is non-convex in the data term  $\lambda(\mathbf{x})C(\mathbf{x}, \rho(\mathbf{x}))$  and convex in regularizer term  $g(\mathbf{x}) \|\nabla \rho(\mathbf{x})\|_{\epsilon}$ . To find a global local minimum, we approximate the energy function with an auxiliary map  $a : \Omega \rightarrow \mathbb{R}$  used to couple the two terms, as done in Newcombe et al., 2011a and Concha et al., 2015:

$$E(\rho, a) = \int_{\Omega} \left\{ \lambda(\mathbf{x})C(\mathbf{x}, a(\mathbf{x})) + \frac{1}{(2\theta)}(\rho(\mathbf{x}) - a(\mathbf{x}))^2 + R(\mathbf{x}, \rho(\mathbf{x})) \right\} d\mathbf{x} \quad (5.9)$$

The coupling term  $\frac{1}{(2\theta)}(\rho(\mathbf{x}) - a(\mathbf{x}))^2$  enforces  $\rho(\mathbf{x})$  and  $a(\mathbf{x})$  to be equal as  $\theta \rightarrow 0$ , at which point  $E(\rho, a = 0) = E(\rho)$ . The global minimum of the convex term  $\frac{1}{(2\theta)}(\rho(\mathbf{x}) - a(\mathbf{x}))^2 + R(\mathbf{x}, \rho(\mathbf{x}))$  is iteratively computed using primal-dual algorithm Aujol, 2009; Chambolle et al., 2011. At each iteration, given a solution for  $\rho(\mathbf{x})$ , the global minimum of the non-convex-term  $\lambda(\mathbf{x})C(\mathbf{x}, a(\mathbf{x})) + \frac{1}{(2\theta)}(\rho(\mathbf{x}) - a(\mathbf{x}))^2$  is found by performing an exhaustive search on  $a(\mathbf{x})$  among the set of  $\xi$  discrete values covering the range of inverse depths  $[\rho_{min}, \rho_{max}]$ :

$$\arg \min_{a(\mathbf{x})} \lambda(\mathbf{x})C(\mathbf{x}, a(\mathbf{x})) + \frac{1}{(2\theta)}(\rho(\mathbf{x}) - a(\mathbf{x}))^2 \quad (5.10)$$

#### 5.4.5.1 Solution

We detail the iterative solution of the energy function. Using vector notation, the convex term is replaced by its conjugate in the primal-dual form using Legendre-Fenchel transform (details and proofs can be found in Aujol, 2009; Chambolle et al., 2011; Handa et al., 2011):

$$\arg \max_{q, \|q\|_2 \leq 1} \left\{ \langle gA\rho, q \rangle - \delta_q(q) - \frac{\epsilon}{2} \|q\|_2^2 \right\} \quad (5.11)$$

where  $q$  is the dual variable,  $A\rho$  computes the  $2MN \times 1$  gradient vector of  $\rho$ ,  $g$  is element-wise weighting defined in eq. (5.8),  $\langle \cdot \rangle$  is the dot product and  $\delta_q$  is an indicator function (Handa et al., 2011) such that for each element  $q$ ,

$$\delta_q(q) = \begin{cases} 0 & \text{if } \|q\|_1 \leq 1 \\ \infty & \text{otherwise} \end{cases} \quad (5.12)$$

Replacing the regularizer with the dual form, the primal variable  $\rho$  and dual variable  $q$  are coupled with the data term giving the sum of convex and non-convex functions to minimize:

$$\arg \max_{q, \|q\|_2 \leq 1} \left\{ \arg \min_{\rho, a} E(\rho, a, q) \right\} \quad (5.13)$$

$$E(\rho, a, q) = \left\{ \langle gA\rho, q \rangle - \delta_q(q) - \frac{\epsilon}{2} \|q\|_2^2 + \lambda C(a) + \frac{1}{2\theta} (\rho - a)^2 \right\} \quad (5.14)$$

Assume a fixed value for  $a$ , the condition of optimality is met when  $\partial_{\rho, q} E(\rho, a, q) = 0$ . Hence, the differentiation with respect to dual variable  $q$ ,

$$\frac{\partial E(\rho, a, q)}{\partial q} = gA\rho - \epsilon \in q \quad (5.15)$$

In case of the primal variable,  $\rho$  the differentiation will be:

$$\frac{\partial E(\rho, a, q)}{\partial \rho} = gA^T q + \frac{1}{\theta} (\rho - a) \quad (5.16)$$

where  $\langle gA\rho, q \rangle = \langle gA^T q, \rho \rangle$  from the divergence theorem, where  $A^T$  forms the negative divergence operator. The complete optimization is solved iteratively, starting at iteration  $t = 1$  where  $\theta$  initialized at  $\theta^1$ . Both  $\rho(\mathbf{x})$  and  $a(\mathbf{x})$  are initialized with the initial depth map obtained from Section 5.4.4, iterating:

1. For the dual variable  $q$  we perform a gradient ascent step and a descent step for  $\rho$ , where the energy has to be maximized for the first and minimized for the second, resulting in the following update step after rearranging terms:

$$\begin{aligned} \frac{q^{n+1} - q^n}{\sigma_q} &= gA\rho^n - \epsilon \in q^{n+1} \\ q^{n+1} &= (q^n + \sigma_q gA\rho^n) / (1 + \sigma_q \epsilon) \\ q^{n+1} &= q^{n+1} / \max(1, |q^{n+1}|) \end{aligned}$$

$$\begin{aligned} \frac{\rho^{n+1} - \rho^n}{\sigma_\rho} &= -gA^T q^{n+1} - \frac{1}{\theta^n} (\rho^{n+1} - a^n) \\ \rho^{n+1} &= \left( \rho^n + \sigma_\rho \left( -gA^T q^{n+1} + \frac{a^n}{\theta} \right) \right) / \left( 1 + \frac{\sigma_\rho}{\theta} \right) \end{aligned}$$

where  $\sigma_q$  and  $\sigma_\rho$  are free parameters used for the differentiation step.

2. At each  $\rho^{n+1}$ , perform a point-wise exhaustive search for minimizing eq. (5.10).
3. if  $\theta^{(t+1)} > \theta_{end}$  got to step 1, otherwise end, where  $\theta^{(t+1)} = \theta^t(1 - \kappa t)$ .

The accuracy depends on the discretization level used for the cost volume construction. To obtain a sub pixel accuracy, we perform a single Newton step proposed by Newcombe et al., 2011a at each iteration.

## 5.5 Live alignment of keyframe depth maps

To obtain a global and consistent reconstruction, we align the computed depth maps with the discrete SLAM map in a single coordinate frame. Most sparse SLAM points have a corresponding 3D point in the dense maps, and we use these as anchors. The anchors are used to keep depth maps aligned with the sparse SLAM map, so that any update in the SLAM map leads to a realignment of the dense maps.

Recall that after each SLAM BA, both the sparse points and the keyframe poses are refined. This refinement may produce a misalignment of the dense maps with respect to the SLAM map. This refinement may not only involve rotation and translation but also a scale change. For this reason, we propose to align each depth map with a similarity transformation. For depth map computed from keyframe  $I_r$ , we perform a non-linear minimization of the reprojection error of visible sparse SLAM points  $\mathcal{P}_L \subset \mathcal{P}$  in  $I_r$  to estimate the similarity transform  $\mathbf{S} \in \mathbf{Sim}(3)$ , using the neighboring keyframes  $\mathcal{K}_L \subset \mathcal{K}$  that share the most feature points with  $I_r$ :

$$\arg \min_{\mathbf{S} \in \mathbf{Sim}(3)} \sum_{i \in \mathcal{K}_L, j \in \mathcal{P}_L} \rho_h (\|\mathbf{x}_{i,j} - \pi(\mathbf{T}_i, \mathbf{S}\mathbf{X}_j)\|^2) \quad (5.17)$$

where  $\mathbf{x}_{i,j}$  is the image observation of the sparse SLAM point  $j$  in keyframe  $i$  and  $\mathbf{X}_j$  is its 3D location from the dense map of keyframe  $I_r$ , in reconstruction coordinates. eq. (5.17) is repeated for every keyframe to align its corresponding depth map.

## 5.6 Experimental Results

### 5.6.1 Benchmark hardware and compared methods

The proposed system has been implemented in C++ and OpenCV using a commodity desktop computer 8GB RAM and GeForce GTX 680 GPU with an Intel(R) Core i7 CPU 3.4GHz. We provide a quantitative evaluation of the reconstruction accuracy with respect to a leading stereo methods (cf. Section 5.6.3). Furthermore, we evaluate the proposed system with the closest dense SLAM method: LSD-SLAM Engel et al., 2014 (cf. Section 5.6.3). For comparison, we tune LSD-SLAM differently on each dataset to achieve best results. LSD-SLAM minimizes SSD of the residual photometric errors assuming rigid image alignment for real-time camera tracking, thus suffers from several drift due to small respiration deformation and severe lighting changes. These minor tracking drifts, typically, corrupt the resultant reconstruction and leads to duplication of points. We also compare with a state-of-the-art dense SFM method Langguth et al., 2016 (cf. Section 5.6.3.5). More details can be appreciated in our video <sup>1</sup>.

### 5.6.2 Datasets

The heart phantom dataset in Mountney et al., 2010b; Stoyanov et al., 2010 is commonly used for dense reconstruction accuracy assessment in laparoscopy because there is ground truth available. Unfortunately, we cannot use it for the evaluation because the camera pose is fixed along the sequence, so it is not possible to compute the 3D reconstruction with any SLAM/SfM

---

<sup>1</sup><https://www.youtube.com/watch?v=RJCmUY9hBSQ&feature=youtu.be>

method. We used several exploratory sequences from public Mountney et al., 2010b and new private datasets with camera motion, recorded by a stereo-laparoscope during the evaluation. The ethical approval for animal use is indicated in Appendix A. Figure 5.3(a-f), shows the typical frames of the evaluation sequences. Figure 5.3(a,b,e) corresponds to sequences of live pigs with strong (cf. Figure 5.3(a)) or small (cf. Figure 5.3(b,e)) respiration. Figure 5.3(c,d,f) corresponds to ex-vivo sequences. The evaluation sequences had different complexities such as weak textures (cf. Figure 5.3(b)(e)) and repetitive textures (cf. Figure 5.3(a)(c)(d)(f)) with either smooth or strongly curved surfaces. The length of the sequences ranged between 20 seconds to 8 minutes.

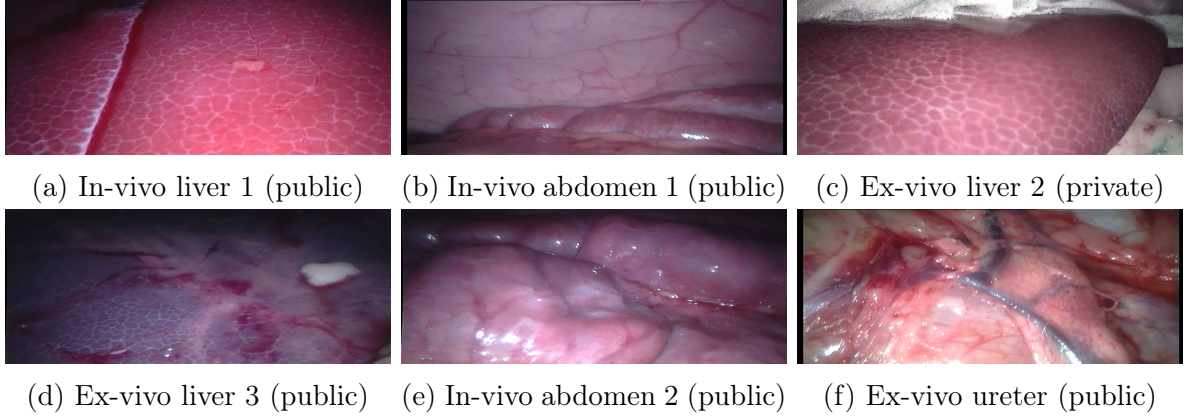


FIGURE 5.3: Sample frames of the different laparoscope porcine sequences used from public Mountney et al., 2010b and private datasets.

### 5.6.3 Quantitative evaluation using dense stereo

Our system was used to reconstruct the scene using only images from the left laparoscope camera. To evaluate, we used their associated right images, and obtained a dense stereo reconstruction using two leading methods Hirschmuller, 2008 and Chang et al., 2013 as our gold standard. According to Maier-Hein et al., 2014b, the stereo method of Chang et al., 2013 is a top performing method for endoscopic images. Figure 5.4 shows the two cameras of the stereo-laparoscope with a red line connecting their optical centers, with our monocular cluster of frames shown in grey. Our reconstruction is up to scale (as with any monocular method), thus before the comparison we estimate the monocular scale factor,  $\mathfrak{s}$ , by means of Least Median of Squares:

$$\arg \min_{\mathfrak{s}} \operatorname{median}_i \|\mathbf{D}_{S_i} - \mathfrak{s}\mathbf{D}_{M_i}\|^2 \quad (5.18)$$

where  $\mathbf{D}_{M_i}$  and  $\mathbf{D}_{S_i}$  are two depths of pixel  $i$  reconstructed by our monocular system and the stereo method, respectively.  $\mathfrak{s}$  is estimated only once for the whole reconstruction of each sequence. The stereo reconstruction is shown with the pixel intensities in Figure 5.4, while the scaled monocular reconstruction is in green. The Euclidean distances between all pixels from the two reconstructions are visualized in Figure 5.5.

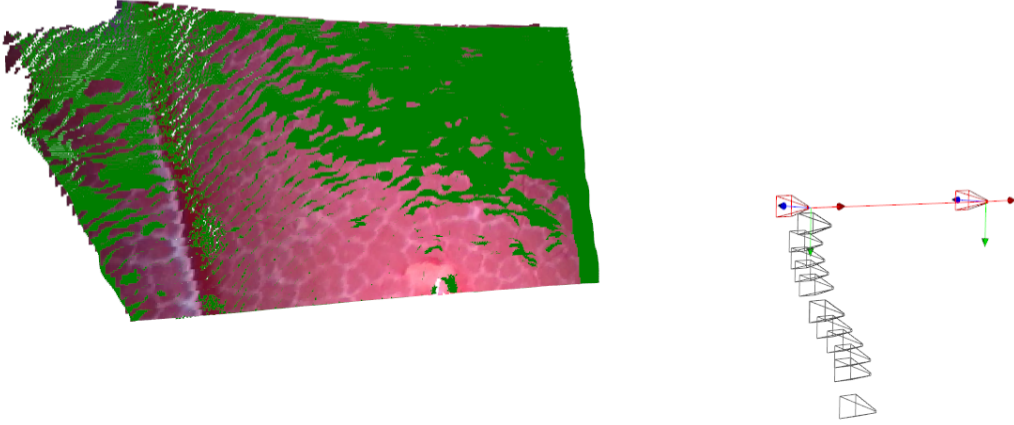


FIGURE 5.4: Monocular (green) and stereo (textured) reconstruction after scale alignment. Stereo cameras are show in red, and cluster of frames in grey.

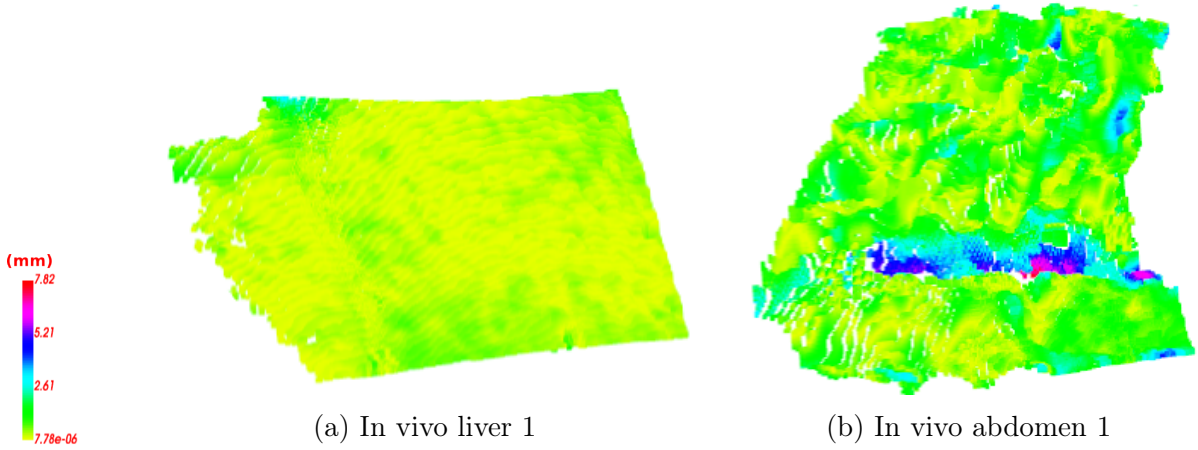


FIGURE 5.5: Euclidean distances between the stereo and the monocular dense maps.

### 5.6.3.1 Evaluation metrics

Table 5.1 reports the averaged reconstruction error. The *reconstruction density per keyframe* is the percentage of pixels reconstructed per keyframe. A reconstructed pixel is one that is visible in all the frames of the cluster, not deleted as a specularity and not located outside the laparoscope’s optical ring. The *stereo coverage* metric is the percentage of monocular reconstructed pixels for which the stereo method computed its depth. For each reconstructed pixel we computed the parallax rendered by the extreme frames of the cluster. Table 5.1 column 5 reports the average parallax among all the reconstructed pixels in all keyframes. We also report the average parallax rendered by the stereo algorithm in Table 5.1 column 6. Per each sequence we tune our monocular algorithm with different parallaxes by means of  $\alpha_1, \alpha_2$ . The RMSE metric is computed as follows. We took all pixels in all keyframes for which both our method and the stereo method computed a depth estimate and measured the distance in the estimated depths. We did this with respect to both stereo methods of Hirschmuller, 2008

and Chang et al., 2013. Table 5.1 also reports average reconstruction density and average reconstruction error of LSD-SLAM except for liver 1 sequence because it has failed due to the strong respiration.

TABLE 5.1: Average reconstruction error with respect to stereo methods (Hirschmuller, 2008, Chang et al., 2013).

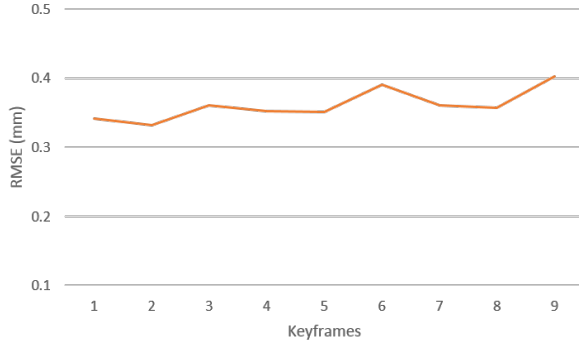
Sequence	Method	Reconst. density per keyFrame %	Stereo coverage %	Mono plx (deg)	Stereo plx (deg)	Avg. RMSE (mm) (Hirschmuller, 2008)	Avg. RMSE (mm) (Chang et al., 2013)
In-vivo liver 1	Proposed system	65	89	0.4	13.1	2.6	2.8
		66	90	5.2		1.0	1.2
		56	90	12.3		0.3	0.4
	LSD-SLAM	X	X	X		X	X
In-vivo abdomen 1	Proposed system	66	79	1.4	8.9	4.5	4.7
		47	88	6.1		2.9	3.3
		32	86	10.1		1.2	1.7
	LSD-SLAM	1.1	98	-		5.4	6.1
Ex-vivo liver 2	Proposed system	48	88	9.1	12	0.8	1.1
		44	79	14.9		0.7	0.9
	LSD-SLAM	1.6	85	-		2.1	2.6
Ex-vivo liver 3	Proposed system	35	98	9.8	11.4	0.5	0.7
		27	98	14.5		0.4	0.5
	LSD-SLAM	3	76	-		1.7	2.4
In-vivo abdomen 2	Proposed system	65	84	2.3	9.6	3.5	3.9
		45	95	4.8		2.9	3.3
		33	92	10.1		1.9	2.2
	LSD-SLAM	2.1	98	-		4.1	5.3
Ex-vivo ureter	Proposed system	58	82	2.9	8.5	2.0	2.3
		45	88	6.0		1.5	2.2
		43	90	11.7		1.0	1.9
	LSD-SLAM	1.4	92	-		2.9	3.7

### 5.6.3.2 Results analysis

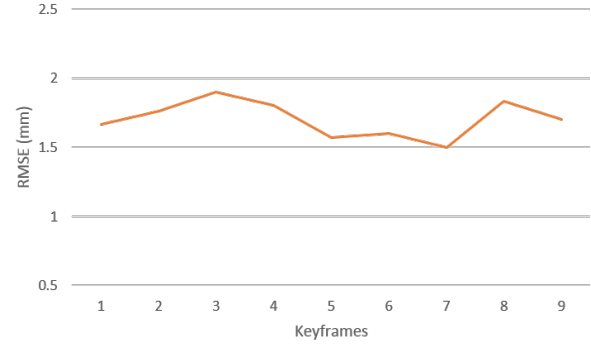
In the same or higher monocular parallax cases with respect to the stereo methods we achieve  $\leq 1.2 \pm 0.8$  RMSE. In such cases, it is difficult to identify whether the remaining error comes from the monocular or the stereo reconstruction. Figure 5.6 shows the RMSE error evolution between keyframes selected by our system to compute the dense reconstruction, in higher parallax case for each sequence. It is important to note that the error between selected keyframes is not incremental, thanks to the careful selection of the cluster of the neighbored images that maintain a minimum parallax irregardless the motion of the camera.

In low parallax cases, the RMSE is higher because there is more corrections by the regularizer (cf. Sec(5.6.3.3)). Table 5.1 also shows a superior performance of the proposed system compared to LSD-SLAM in terms of reconstruction density and accuracy. Figure 5.7, shows the reconstruction of our system and LSD-SLAM from different points of view. For in-vivo

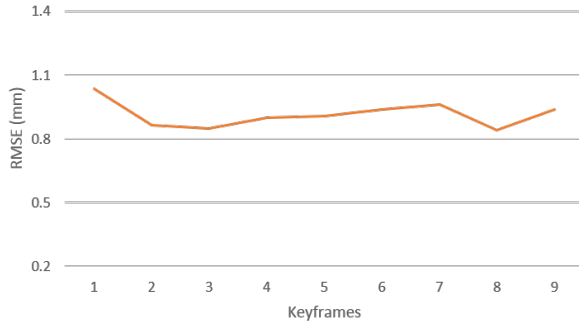
sequences the proposed method was robust to small respiration deformation as inter-frame motion in the frames cluster was considerably small. Live incremental reconstruction results can be seen in our video <sup>2</sup>



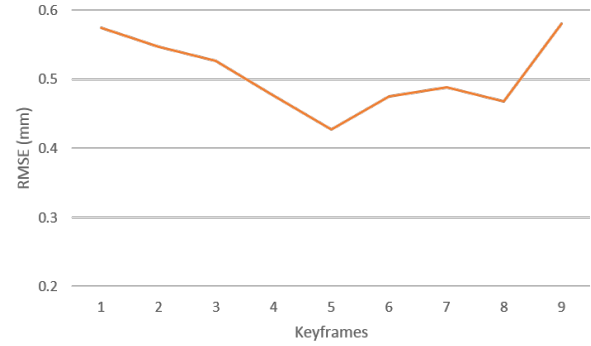
(a) In-vivo liver 1.



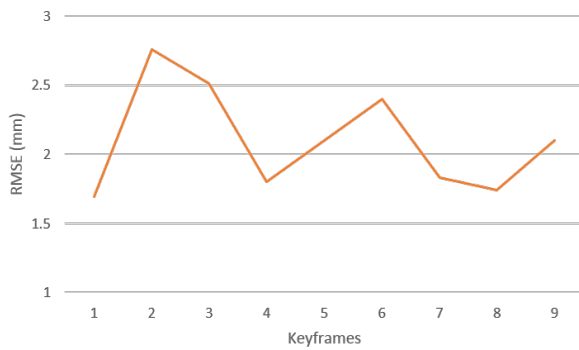
(b) In-vivo abdomen 1.



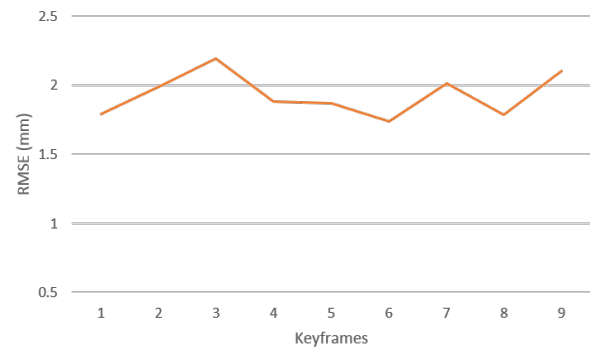
(c) Ex-vivo liver 2.



(d) Ex-vivo liver 3.



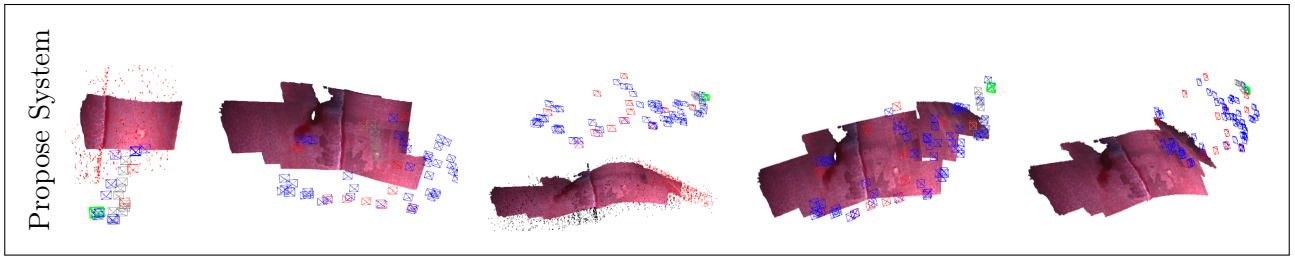
(e) In-vivo abdomen 2.



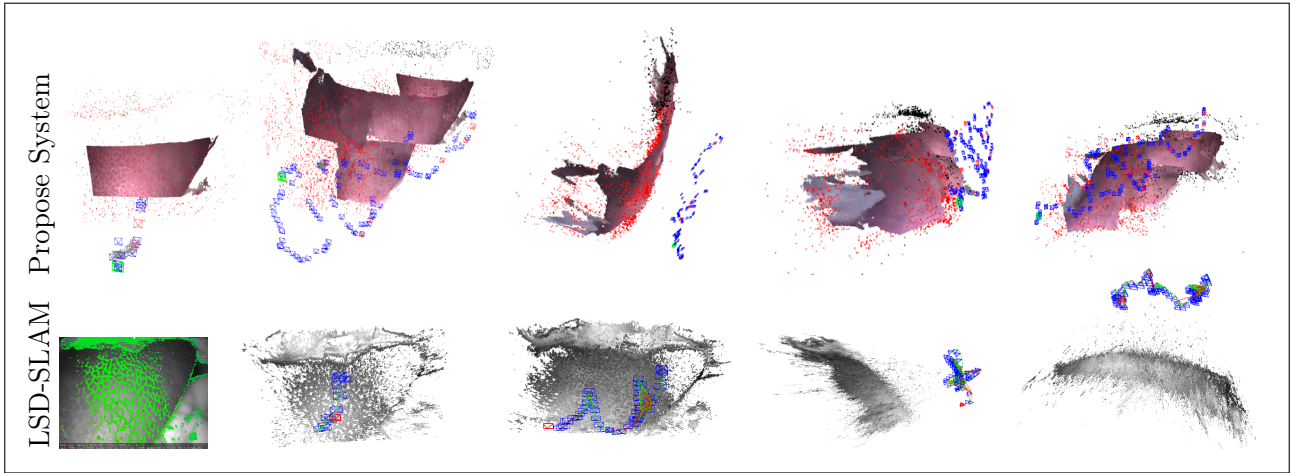
(f) Ex-vivo ureter.

FIGURE 5.6: Error evolution between selected keyframes for dense reconstruction.

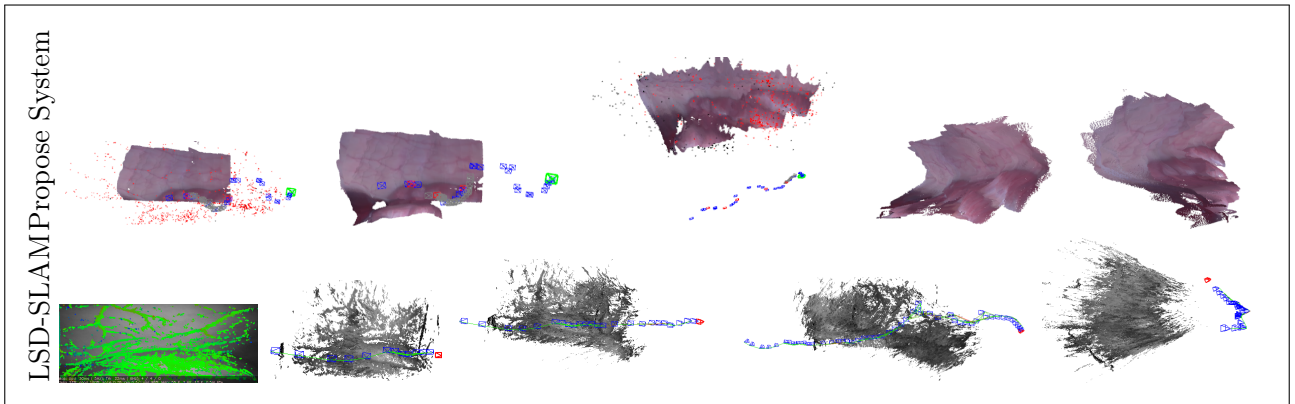
<sup>2</sup><https://www.youtube.com/watch?v=RJCmUY9hBSQ&feature=youtu.be>



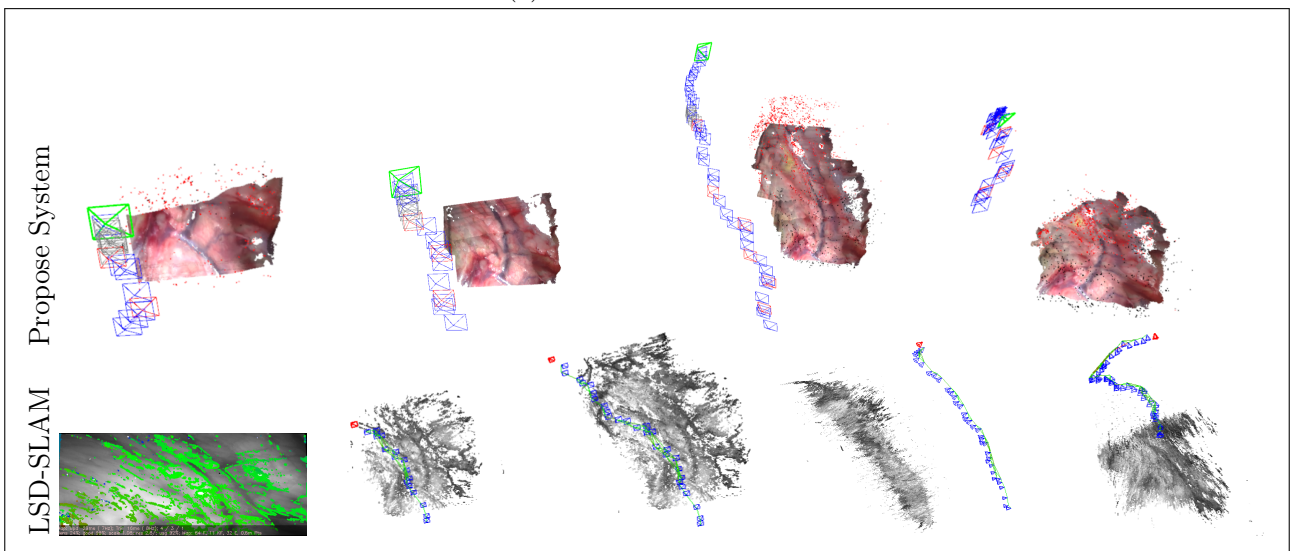
(a) In-vivo liver 1



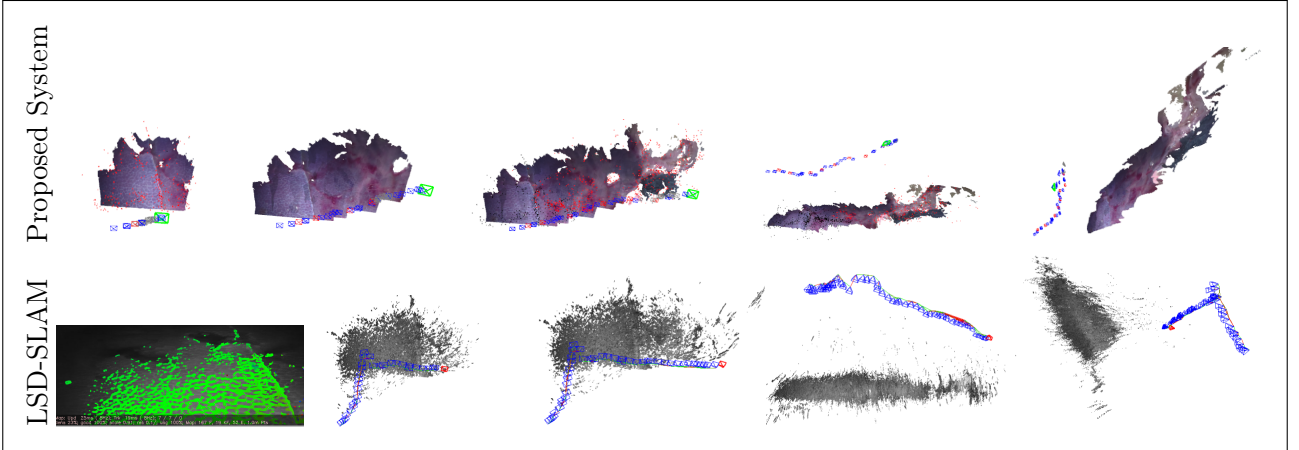
(b) Ex-vivo liver 2



(c) In-vivo abdomen 1



(d) Ex-vivo ureter



(e) Ex-vivo liver 3

FIGURE 5.7: Incremental dense reconstruction of proposed system and LSD-SLAM on different sequences visualized as point clouds. SLAM keyframes and points are colored in blue and in red, respectively. The selected keyframes used for the dense reconstruction and frames cluster are colored in red and grey, respectively. The green frustum shows the current laparoscope pose.

### 5.6.3.3 The influence of the regularizer and number of images in the cluster

We analyzed the effect of the regularizer in low parallax cases in Figure 5.8(a,d). It shows how the correction made by the regularizer in the variational optimization is proportionally bigger in low parallax cases. It can be seen also how the RMSE is smaller in the case of the liver than in the abdomen. We conjecture that it is due to the fact that the liver surface geometry is smoother than that of the abdomen, and hence fits better the regularizer prior, which favor smooth reconstruction and because of that the final error is smaller. In high parallax cases, Figure 5.8(b,e), the regularizer effect is minimal, and its effect is to remove the stair-casing effect and provide a smoother reconstruction.

The quality of the reconstruction is mostly dependent on the data term, and increasing the number of cluster images generally improve the accuracy. In Figure 5.8(b,c) and (e,f) we show a comparison of the reconstruction obtained using all the frames in the cluster vs. using only the two extreme frames in the cluster. The data term is a simple two view stereo when using two images, and the lack of data constraints can lead to spurious local minimum in the variational problem. However, the cost when using a cluster of many images taken from different viewpoints generally produces a strongly constraint problem with a strong global minimum. This directly increases the chance that a good initial solution is found (cf. Section (5.4.4)).

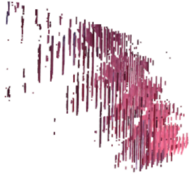



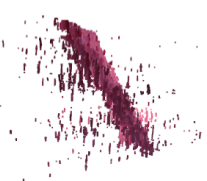


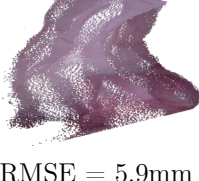

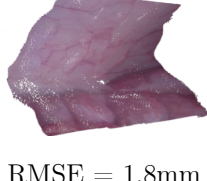
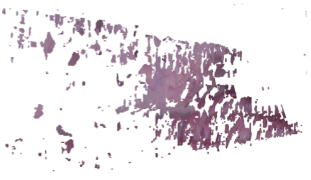
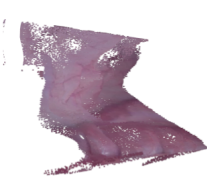
	Pllx.	Num. of used Images	Initial Reconstruction (Section 5.4.4)	Regularized
In-vivo liver 1	0.37°	All images in cluster	 RMSE = 40.0mm	 RMSE = 2.8mm
			(a)	
	12.3°	All images in cluster	 RMSE = 1.6mm	 RMSE = 0.4mm
			(b)	
	12.3°	Two extreme images in cluster	 RMSE = 16.8mm	 RMSE = 2.5mm
			(c)	
In-vivo abdomen 1	1.4°	All images in cluster	 RMSE = 85.7mm	 RMSE = 5.9mm
			(d)	
	10.1°	All images in cluster	 RMSE = 7.9mm	 RMSE = 1.8mm
			(e)	
	10.1°	Two extreme images in cluster	 RMSE = 65.6mm	 RMSE = 8.6mm
			(f)	

FIGURE 5.8: Effect of the regularizer and the number of processed images in the cluster.

#### 5.6.3.4 Robustness of ZNCC versus SAD for the data term

Due to lack of open source implementation of DTAM, we couldn't make a comparison with DTAM. Besides the dense tracking functionality of DTAM versus the sparse tracking of our system, another major difference between both systems is the data term being used. Thus, here we illustrate the effect of ZNCC as data term in comparison to SAD used by other dense SLAM systems Newcombe et al., 2010; Newcombe et al., 2011a; Concha et al., 2015.

Figure 5.9 shows two depth maps obtained by both data terms for images shown in Figure 5.3(a,b), with 13 frames in the cluster and high parallax. Despite the ZNCC data term being more computationally expensive than SAD, it provides a better initial depth map estimation. Thus, with ZNCC, it reduces the regularizer effort to mainly smoothing the solution.

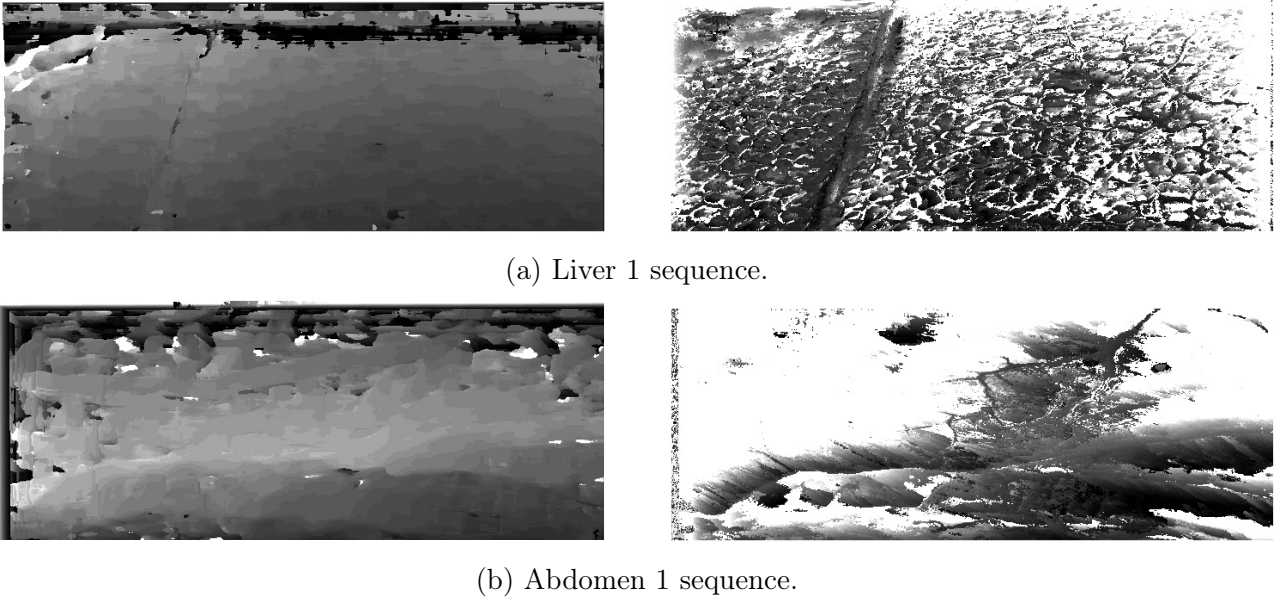


FIGURE 5.9: Initial depth maps obtained when using ZNCC (left) and SAD (right) for the data term in the variational problem.

Typically, errors exit for pixels at the margins of the image because they are non-visible in all frames in the cluster. Table 5.2 reports the RMSE errors for the initial reconstruction extracted from the cost volume when using ZNCC and SAD in eq. (5.6), in addition to the final errors after variational minimization.

TABLE 5.2: RMSE errors in mm with respect to Chang et al., 2013, when use ZNCC and SAD for the data term in the variational problem.

	ZNCC		SAD	
	Initial Reconstruction	Energy Minimization	Initial Reconstruction	Energy Minimization
Figure 5.9 (a)	1.1	0.26	68.4	38.3
Figure 5.9 (b)	6.4	1.8	118.3	101.7

### 5.6.3.5 Proposed system versus dense SfM

We evaluated the reconstruction accuracy and computation time with a state-of-the-art dense SfM method Langguth et al., 2016. We have performed this evaluation on the ex-vivo liver 3 sequence. Figure 5.10(a,b) shows the final reconstruction by Langguth et al., 2016 after the filtering/refinement step. The RMSE was 0.6mm and 0.8mm with respect to stereo methods Hirschmuller, 2008; Chang et al., 2013, respectively. The averaged rendered parallax was  $12.4^\circ$ . The proposed system and Langguth et al., 2016 yields similar accuracy and both render higher monocular parallax than stereo methods, however the proposed system is order of magnitude faster. The dense reconstruction of Langguth et al., 2016 took  $\approx 4.5$  min and the subsequent filtering step took  $\approx 1.5$  min.

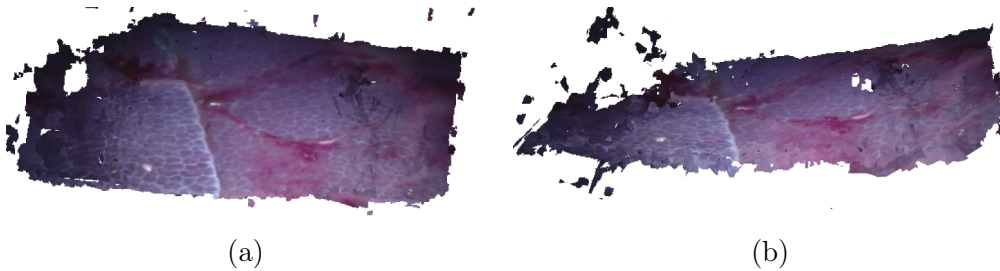


FIGURE 5.10: Dense SfM (Langguth et al., 2016).

### 5.6.4 Qualitative evaluation on patient data

The proposed system has been qualitatively evaluated on a short exploratory sequence for the abdominal cavity of one patient. The sequence has been recorded by a surgeon who has performed a challenging laparoscope exploration with fast movements and orientations changes, without prior knowledge or guidance about SLAM. Figure 5.11(a) shows image sample of patient liver sequence, as can be seen the liver textures are very challenging and far less than in pig liver. However, our SLAM was able to locate few but accurate features points to robustly estimate laparoscope camera poses. We show in Figure 5.11(b) the sparse SLAM reconstruction. Figure 5.11(c-d) shows our dense reconstruction results, more details can be seen in our video <sup>3</sup>.

---

<sup>3</sup><https://youtu.be/GrA30U6t8KE>

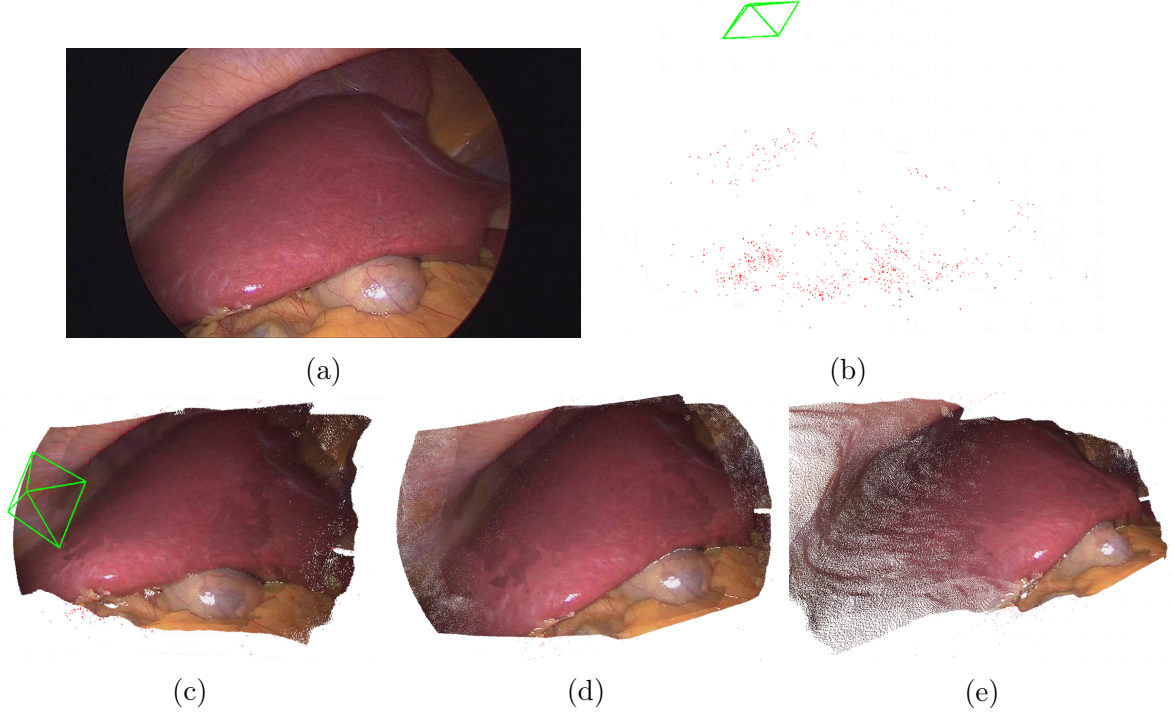


FIGURE 5.11: Reconstruction results on patient liver sequence. (a) Image sample, (b) Sparse ORB-SLAM reconstruction. (c-d) Dense reconstruction of liver surface by our system, from different directions.

### 5.6.5 Free parameters tuning

We detail in Table 5.3 all the free parameters which were used with the same tuning for all the experiments except  $\beta_{max}$ , where it is set to 10 with abdomen 1 sequence only, because the sparse points were very few and not describing the scene depths properly. The most sensitive parameter is the overlap between the cluster frames, controlled by  $\alpha_1$  and  $\alpha_2$ . We fix  $\varkappa = 0.001$  to meet a good balance between the quality and computing time trade-off in the variational minimization. Integral images were used to keep the running time invariant to the ZNCC window size as proposed in Stoyanov et al., 2010.

TABLE 5.3: Parameters tuning.

$\alpha_1$	$\alpha_2$	$\theta^1$	$\varkappa$	$\theta_{end}$	$\lambda$	$\omega$	$\mathcal{W}$	$\beta_{min}$	$\beta_{max}$	$\xi$	$\epsilon$	$\tau$	$\psi$	$\sigma_q$	$\sigma_\rho$
0.2	0.01	0.2	0.001	0.0005	0.5	0.01	19	0.8	5	51	0.001	30	0.2	20	0.5

### 5.6.6 Processing time

We report in Table 5.4 the average execution time needed by each step of the proposed system, for dense reconstruction, and the average execution time of the two parallel threads from ORB-SLAM (Sparse Tracking and Sparse Reconstruction) for different image resolutions.

TABLE 5.4: Average processing time (in seconds).

Image Resolution	Sparse Tracking	Sparse Reconstruction	Dense Reconstruction					
			Cluster selection	BA	Inverse depth Discretization	Cost volume	Variational minimization	Depth maps realignment
720x288	0.03	0.60	0.17	1.3	0.00036	3.4	6.2	0.38
960x260	0.04	0.69	0.21	2.0	0.0039	5.2	8.4	0.47

In the *Dense Reconstruction* thread for image resolution 720x288 of public dataset, the selection of the reference keyframe and its frames cluster took  $\approx 0.17$ s followed by a Bundle Adjustment, that accurately estimates the poses of frames in the cluster  $\approx 1.3$ s. It is worth noting that most of this time is spent computing the sparse matches between the frames in the cluster, the BA stage just took  $\approx 100$ ms. The ZNCC cost volume construction took  $\approx 3.4$ s implemented on the GPU and the cluster size varied between 5-18 frames. The equivalent time using CPU implementation varied between 18-25s. The variational solver was implemented on the CPU, yielding a computation time of  $\approx 6.2$ s. Using a GPU implementation as proposed in Newcombe et al., 2011a could reduce this time significantly. The depth maps re-alignment stage took  $\approx 380$ ms on average. In case of our private dataset that has 960x260 image resolution, the processing time are slightly increased due to large number of images features.

### 5.6.7 Augmented reality annotations

The recovered dense geometry of tissue makes AR labeling in surgical scene simpler, where image features are no longer required for anchoring virtual marks. For example, the practitioner/surgeon can add annotations/marks to the location of his/her interest on the dense reconstruction. Figure 5.12(a,b) shows the placement of different annotations on the reconstructed liver surface during laparoscope exploration, from top and lateral view. The recovered surface in Figure 5.12(a,b) has been reconstructed during left-to-right laparoscope motion. Once annotations are added to the virtual scene the relative laparoscope pose estimated by SLAM is used to place a virtual camera, red frustum in Figure 5.12(a,b), in the virtual scene, similar to 3.4.3, to render virtual images to be fused with the input laparoscope images in real time and obtain AR overlay (c.f. 5.12(c)). In order to avoid the jittering effect, the annotation are anchored to dense surface such that any update in the dense surface (c.f. Section 5.5) leads to the same update in the AR annotations. Figure 5.12(d-f) shows the AR annotation during the rest of the sequence where the endoscope was moving in the reverse direction, right-to-left. More details can be seen on our video <sup>4</sup>

<sup>4</sup><https://www.youtube.com/watch?v=RJCmUY9hBSQ&feature=youtu.be>

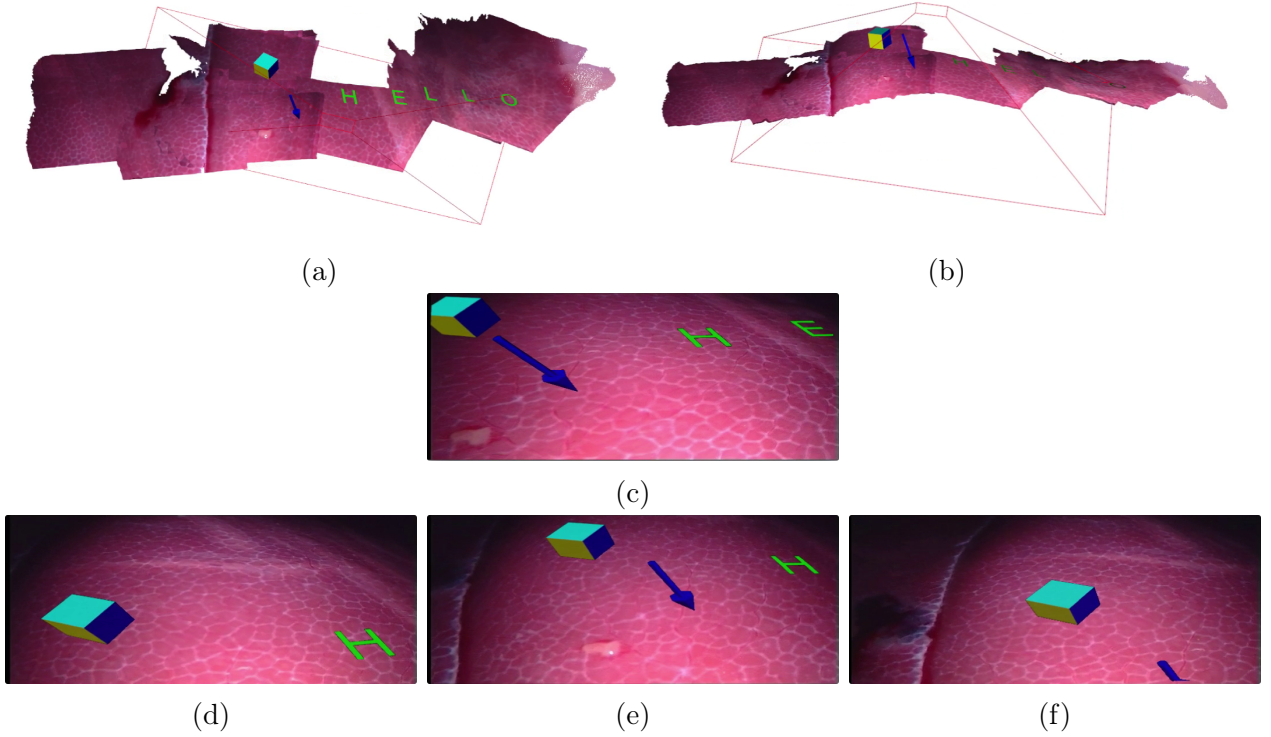


FIGURE 5.12: Using dense surface for AR annotations. (a,b) Adding AR annotation on the reconstructed dense surface, red frustum is virtual camera placed at estimated laparoscope pose. (c) AR annotation at estimated laparoscope pose in (a,b). (d-f) AR view during laparoscope exploration in right-to-left direction

### 5.6.8 Performance on indoor sequences

The proposed system has been quantitatively evaluated on different indoor sequences from TUM public RGB-D dataset (Sturm et al., 2012) of a moving monocular camera. We show our dense reconstruction results on a subset of the sequences where most RGB-D methods are usually use. To do so, few system parameters were slightly changed, where we set  $\omega = 0.1$  to downweight the regularizer strength, where ZNCC data term of window size  $\mathcal{W} = 11$  have proven to be reliable in such environments. We reduced  $\alpha_2$  to 0.005 to include more frames in the cluster and thus ensure a high quality matches from small baseline frames.

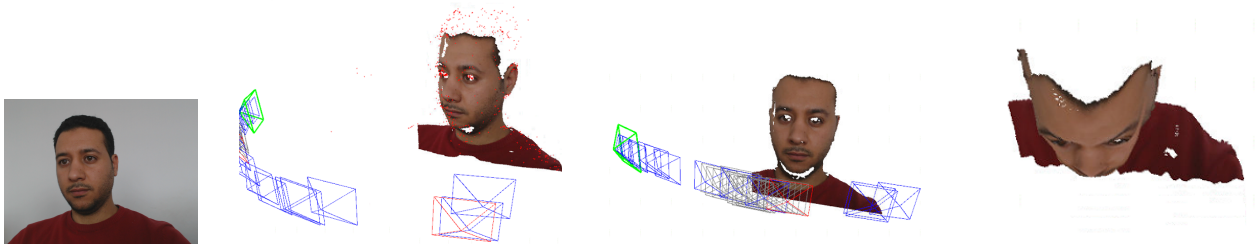
We show in Figure 5.13 the incremental dense reconstruction of the scene during camera exploration from one or several meters distance. Figure 5.13(a) shows the dense reconstruction of the teddy bear that has a soft fur and wears a yellow smooth shirt in addition to the Euclidean distances of the reconstructed pixel from the given image sample on the left. Similarly, we show in Figure 5.13(b) the dense reconstruction of a planar zig-zag (orthogonal) structure together with the Euclidean distances of the reconstructed pixel. The system has also been tested for face reconstruction from the same sequence used in Section 4.6.6, of a sitting person with a camera moving around his face (cf. Figure 5.13(c)). Figure 5.13(d) shows the reconstruction from a moving camera, that was one meter high and moved in circle around a planar surface (several conference posters sticked to the floor), the beginning and the end of the trajectory overlap, so that there is a loop closure. As can be seen the superior reconstruction density of our approach over LSD-SLAM in Figure 5.13(d) right. Additionally, we show the reconstruction of a desktop environment that contains homogenous regions that are very difficult to reconstruct.



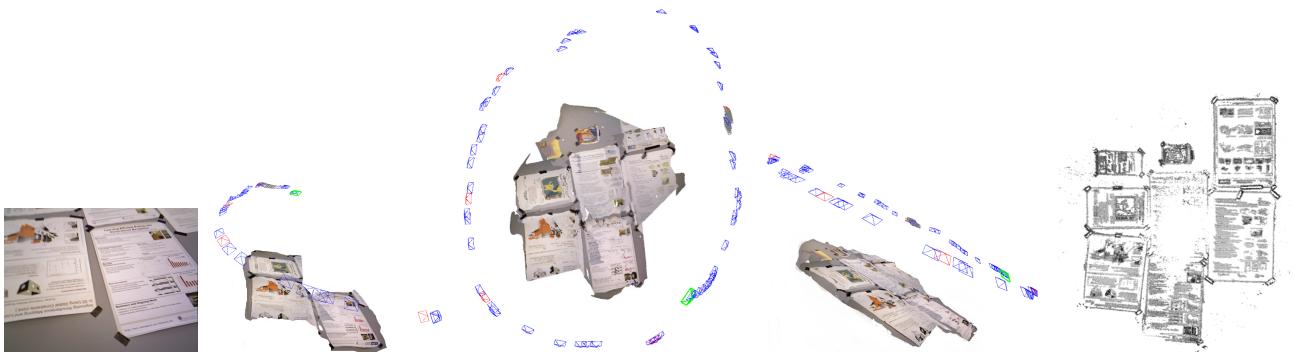
(a) Sequence “*fr3:teddy*”. Right:Euclidean distances with respect to ground truth of image sample on left.



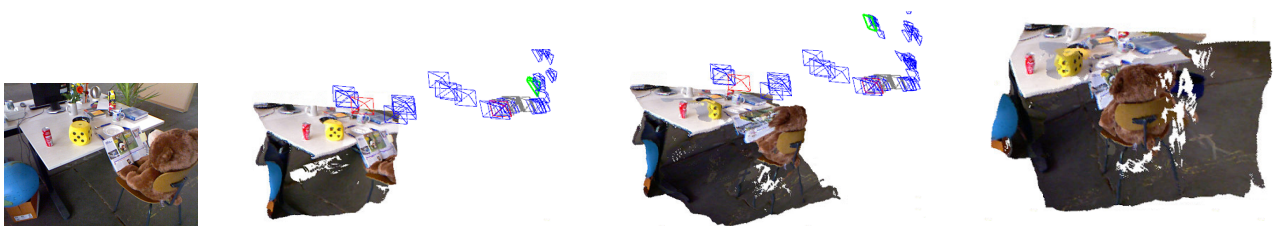
(b) Sequence “*fr3:structure\_texture\_near*”. Right:Euclidean distances with respect to ground truth of image sample on left (top view).



(c) *Face sequence*.



(d) Sequence “*fr3:nostructure\_texture\_near\_withloop*”. Right, LSD-SLAM reconstruction.



(e) Sequence “*fr2:desk*”.

FIGURE 5.13: Performance on different indoor scenes from public dataset (Sturm et al., 2012).

Table 5.5 reports a quantitative evaluation with respect to ground truth information available at TUM dataset (Sturm et al., 2012), except for Face sequence because no ground truth was available. For each processed sequence, our systems selects a set of keyframes to densify the scene, and we computed the RMSE of the Euclidean distances for all pixels reconstructed from those keyframes. The Euclidean distance is computed for the pixel that has depth estimate in the ground truth depth map and our estimated depth map. We provide in Table 5.5 the averaged RMSE, in centimeter, for each processed sequence. Before the comparison, we estimate the monocular scale factor,  $s$ , by means of Least Median of Squares, similar to eq. (5.18). Concha et al., 2015 has reported average RMSE 2.8cm for *fr3:nostructure\_texture\_near\_withloop* sequence, while our RMSE is 3.5cm (no RMSE is reported for the reset of the sequences in Figure 5.13). Concha et al., 2015 assume that homogeneous color regions belong to planar areas and they use that prior in their dense mapping approach, thus having lower error because the scene consists of planar structures. However, without explicit scene prior our system shows a promising accuracy. *fr2:desk* sequence is very challenging, where many difficult homogenous regions exist, e.g: floor, desk, computer screen, ... etc., which causes high RMSE as our system cannot perform well in such homogenous regions. The image resolution of all the sequences is 640x480, and the average processing times needed for: 1) image cluster selection and BA was 1.8 seconds; 2) ZNCC cost volume was 7 seconds; 3) variational minimization was 9 seconds.

Sequence	Average RMSE (cm)
<i>fr3:teddy</i>	2.5
<i>fr3:structure_texture_near</i>	2.8
<i>fr3:nostructure_texture_near_withloop</i>	3.5
<i>fr2:desk</i>	6.8

TABLE 5.5: Average reconstruction error with respect to RGB-D sensor.

## 5.7 Conclusion

A novel real-time dense monocular SLAM system has been presented in this chapter that uses the sole input of frames from a standard monocular endoscope. The proposed system is able to track the endoscope at frame-rate using image features, and benefiting from the acquired SLAM keyframes, it is able to produce a high quality dense reconstruction of the surgical scene. The proposed system has proved to be fast and does not need any external tracking hardware nor intervention at any rate. It therefore can be integrated smoothly into the surgical workflow. Unlike other direct SLAM approaches, the proposed system presented an effective way for neighbored images selection, that are used for scene densification, with local BA to accurately refine their initial estimated poses. Thus, avoids poorly constrained initialization for the dense mapping optimization. Furthermore, the use of illumination invariant image data term achieves a robustness to illumination variability, auto gain or exposure in endoscopy.

It has been validated and evaluated on real in-vivo and ex-vivo laparoscope sequences from public and private datasets and shows a accurate reconstruction with different scene textures. A rigid scene model is assumed, however, the systems has proven a robustness with respect to small deformations resultant from respiration. Unlike pure dense stereo approaches, we can

control the parallax and thus the reconstruction accuracy, which is particularly relevant in cases where the endoscope is relatively far from the scene. In these cases, the stereo endoscope cannot achieve sufficient parallax, due to its fixed baseline. Moreover, the propose system has been qualitatively evaluated on short human sequences and shows very good reconstruction of patient abdominal cavity. Additionally, the proposed dense SLAM system has been evaluated on public indoor RGB-D dataset and showed a very promising reconstruction results, with minor tuning of system parameters. Thus, it can perform equally in endoscopic and indoor scenes.



# Chapter 6

## Conclusions and Future Directions

### Contents

<b>6.1</b>	<b>Conclusions</b>	<b>95</b>
<b>6.2</b>	<b>Future directions</b>	<b>96</b>
6.2.1	Improvements to our dense SLAM system	97
6.2.2	Clinical trials	97
6.2.3	Is dense MIS SLAM the holy grail for AR?	98
6.2.4	Are the rigid assumptions enough?	98
6.2.5	Deep learning	98
6.2.6	Comprehensive and diverse dataset	99
6.2.7	Robotized MIS	99

### 6.1 Conclusions

In this thesis we have devised and validated one of the first monocular SLAM systems able to provide the two vital pieces of information for AR in MIS: *intra-operative dense reconstruction* of the surgical scene and the *relative pose of the endoscope's camera* with respect to the estimated reconstruction. The proposed SLAM does not require any additional inputs other than monocular RGB images of the standard endoscope. The system is close to fit the three ideal requirements stated by Sielhorst et al., 2006: 1) **Usability**: a strict minimum of interaction by surgeon is required; 2) **Interoperability**: generic data and protocols are used, which guarantees the largest compatibility with other equipments within operating room. 3) **Reliability**: the systems provide reliable performance in different situations. Therefore, they system can be smoothly integrated with existing laparoscopic imaging equipment without introducing perturbations to the surgical environment or the clinical workflow.

The system has been validated and evaluated on in-vivo and ex-vivo sequences from public and private datasets. It has been also evaluated on public indoor RGB-D dataset and showed very promising reconstructions. We strongly believe that the proposed system can perform

equally well in other endoscopic sequences coming from flexible endoscopes or even capsule endoscopes.

In Chapter 3, we studied the utilization of visual SLAM as a robust camera estimator and have shown how it is able to provide the key information for reliable on-patient AR visualization using only a Tablet-PC with a built-in camera. It avoids the tedious hardware setups inside the operating room, which wastes too much of the surgeon’s attention. The proposed AR system has showed a high registration accuracy and robustness to scene occlusions, without the need for artificial landmarks on the patient skin for either registration nor camera localization. Next, in Chapter 4, we have researched how feature-based SLAM is able to provide accurate camera tracking in endoscopic sequences. We have shown how with adequate customization, the state-of-the-art monocular ORB-SLAM can provide accurate camera tracking. Thanks to the ORB bag of binary words image recognition, the system was able to robustly relocate the endoscope camera pose after tracking lost, which is one of the important factors for a reliable camera tracker. The endoscope camera can be located very accurately and robustly, however, the map was very poor in terms of density. In a first attempt to densification, we presented a pairwise dense approach that significantly and accurately improved the reconstruction from a sparse set of points to a quasi-dense reconstruction. Generally, wide baseline pairwise matching leads to accurate reconstruction, however it is still “*ill-conditioned*”, inaccuracies in reconstruction are likely to happen, specially in poor textured regions. Lack of data constraints can lead to spurious matches.

In Chapter 5, we devised and validated a system that combines the accuracy of feature based camera localization with a dense approach that not only consider a pair of images but a sequences of very close images with a boost in performance with respect to our previous approach. We presented a novel dense SLAM system, that uses novel dense multi-view stereo-like approach. On one hand, the proposed system can effectively increase the width of the stereo baseline and on the other hand consider more images between the stereo pair, thus better constrain the matching problem and finding a strong global minimum for each pixel. The presented system differs from other direct SLAM approaches in important ways: 1) It uses an efficient criterion for neighbored images selection used for scene densification, with on-the-fly feature based BA that accurately refines their initial estimated poses. 2) Thanks to GPU processing and ZNCC illumination invariant, it is robust to severe illumination variability, auto gain or exposure in endoscopy, additionally it can perform equally on endoscopic and indoor scenes. The system has been proved to be superior to state of the art methods in in-vivo and ex-vivo sequences from public and private datasets. It has been qualitatively evaluated on exploratory sequence of patient and shows very good reconstruction of patient’s internal cavity, however it was very challenging. Additionally, it has been evaluated on public indoor RGB dataset and showed very promising reconstructions.

## 6.2 Future directions

After the initial proof of the advantages of dense monocular SLAM, there are several research venues for future work.

### 6.2.1 Improvements to our dense SLAM system

The presented dense SLAM system combines in a novel way different principles and comes with a real-time tracking and dense reconstruction abilities. However, there are several areas to improve our basic initial proposal:

1. The proposed system cannot deal with very homogeneous soft-tissue surfaces that do not have texture characteristics. Thus, additional visual cue such as shading would improve the reconstruction of these soft regions, that flexibly models the reflectance properties of the surface. That fusion can probably provide a superior reconstruction results than either dense SLAM or SfS can achieve.
2. The system uses image features for camera pose estimation, which have been proven to be sufficient for abdominal cavity interventions. However in other scenarios, such as GI interventions, textural image information tends to be scarce and feature based tracking cannot be so reliable, thus given first limitation solved, one can consider dense based tracking to avoid tracking failure in very homogeneous areas.
3. The proposed system requires offline camera calibration. However, during surgery various camera parameters may be adjusted in order to offer the optimal view for the surgeon such as zoom and focus, which will cause large errors in the reconstruction and the tracking. A technique for detecting these changes and perform online self-calibration is important. Methods to cope with focus changes have been reported by Stoyanov et al., 2005 but have limited capabilities in practical use. An interesting approach to online calibration would be a mix of both pre-calibration and self-calibration as proposed by Pratt et al., 2014.

### 6.2.2 Clinical trials

In the long run, validation of computer assisted systems in clinical trials are needed to fully prove the benefits of these new techniques. There are reports for clinical trials, however they are very limited either in terms of cases or regarding clinical benefit measurement (Bernhardt et al., 2017). Once the benefit of the new approaches has been proven on patient studies, and the new proposed systems are integrated effectively into the clinical workflow, computer assisted surgery systems will find widespread acceptance in clinical routine. Consequently, our future work will consider several clinical validations of the proposed solutions. We believe that the real-time endoscope localization and dense reconstruction system presented in this thesis can have immediate impact on surgical outcomes in terms of operation time, safety, and efficiency. For example in pediatric endoscopic, the laparoscopic splenectomy for hematological disorders diseases is still limited because of the complications in the operating field. Where, in children, spleen size is extremely large for the body size in hematological disorder. So, the working space is very small compared with adult patients. Hence, a SLAM powered navigation system would offer an AR overlay of spleen hidden vascular and pancreatic tail location, in addition to scene dense representation of surrounding anatomical structure to the huge spleen and thus allows for fast, safe and precise endoscopic surgery in children.

### 6.2.3 Is dense MIS SLAM the holy grail for AR?

As already explored in this thesis dense SLAM can provide the key elements needed to achieve a markerless AR system, however an accurate registration method is still a pre-requisite. In this thesis we showed a simple yet way to align the intra-operative models for AR overlay of hidden structures. However, in real scenarios intra-operative image acquisition with proper organ segmentation is a cumbersome and adds to cost, time and not available in conventional clinical workflow. Instead, a registration of pre-operative models is a must, which is a challenging task because organs undergoes different deformations due to gas insufflation at surgery time. To recover such deformation an intra-operative acquisition is used (Bano et al., 2013), that require accurate organ segmentation. Given the presented dense SLAM system that can provide fast and dense scene reconstruction, it would be interesting for the future research to develop a semi-automatic/automatic non-rigid biomedical model, to accurately register the pre-operative data with dense intra-operative reconstruction.

### 6.2.4 Are the rigid assumptions enough?

The fundamental assumption for camera pose estimation in our SLAM is a rigid environment. Although, this holds for abdominal cavity interventions where semi-rigid points are located on diaphragm and abdominal wall and are sufficient for camera pose estimation. This rigid model cannot hold for other surgeries with fully non-rigid tissue motion such as cardiac surgery. A deformable framework must be established for dealing with deformation caused by cardiac motion, organ shift and tissue-tool interaction. Mountney et al., 2010a have modeled the periodic cardiac motion within SLAM architecture. However, complex tissue tool interactions and organ shifts are likely to require complex biomechanical modeling.

We also consider a rigid model for dense reconstruction process. This rigid model is beneficiary to obtain the dense surface information before any surgical interaction. This initial reconstruction is a pre-requisite to track the non-rigid organ deformation during surgical interaction (Bartoli et al., 2015).

### 6.2.5 Deep learning

Deep learning approaches are growing rapidly in both medical and computer vision fields (Twinanda et al., 2017; Tatenno et al., 2017). In the medical field they are limited to surgical phase recognition and tool detection tasks (Twinanda et al., 2017). We believe that deep learning approaches can have a vital role in different aspects for endoscopy, whereas human internals body structures are similar. Hence, a well-trained deep learning approach on this specific environment can help to:

- Detect, describe and match endoscopic image features, which is very important step in different MIS tasks. Currently, for feature detection and description, well-known computer vision approaches are being used extensively, however they are not designed for endoscopic images. These approaches provide a decent performance in few image locations, however they miss many good features that should be considered because they are not fulfilling their predefined priors (e.g: pattern of FAST detector).

- Guide the reconstruction process (e.g: a semantic mapping fashion), thus can improve depth estimation at very homogenous areas. The use of deep learning has demonstrated the potential of regressing depth maps at a relatively high resolution and with a good absolute accuracy even under the absence of monocular cues (motion, texture, shading ...etc.) to drive the depth estimation task (McCormac et al., 2017).
- Estimation of the absolute scale, which is the main limitation of monocular SLAM approaches (Tateno et al., 2017). One advantage of deep learning approaches is that the absolute scale can be learned from examples and thus predicted from a single image without the need of scene-based assumptions or geometric constraints.

However, the performance of deep learning approaches depend strongly on large amounts of annotated training data, which needs many efforts to construct in endoscopy. Nonetheless, our proposed SLAM methods provide geometrical alignment along the image sequences that would be useful to propagate the labelling along the frames of the sequence.

### 6.2.6 Comprehensive and diverse dataset

One issue faced by most image guided surgery systems is the difference between the success criteria from research point of view and those from a medical one. For the researchers, success is mostly demonstrated by the fulfillment of accuracy goals, often demonstrated by numerically simulated data, phantom models with known ground truth geometry, or ground truth data obtained by imaging modalities. For the latter, the success of a system consists in providing real clinical proof of its utility via real parameters such: bleeding, smoke, and tissue cutting. Thus, it is highly needed to have a public and generic dataset to fill the gap between the research criteria and the medical for successful systems. Efforts have been made by Maier-Hein et al., 2014b to assess and compare the accuracy of different dense stereo approaches on a unified dataset with real conditions using a unified evaluation protocols. However, the dataset contains only in-vitro image pairs, thus not suitable for monocular case.

Additionally, having a standard dataset allows researchers to report their validation experiments in a standardized manner with a unique evaluation method. This standardization of the dataset and evaluation procedure enables to keep track of research progress in MIS field and identifying current limitations to move the field forward. The ideal dataset should combine different information (e.g: video sequences, images, and ground truth information) from various and real interventions on both animal and human cases.

### 6.2.7 Robotized MIS

Robotic assistance in MIS, e.g: da Vinci, has gained significant popularity. In that case, the surgeon manipulate the instruments through a remote manipulator that allows him/her to perform the normal movements associated with the surgery, whilst the robotic arms carry out those movements. It incorporates highly dexterous tools, hand tremor filtering, and motion scaling, thus increased the dexterity of surgeon in challenging tasks such as tissue cutting. However, it is still places burden on surgeons to define the optimal motion, e.g: cutting trajectories. Recently, autonomous robotic system have shown higher accuracy than expert human surgeons performing the same tasks, as experimentally proved by semi-autonomous STAR system

(Opfermann et al., 2017). Having a dense tissue reconstruction would boost the development of fully autonomous robot in MIS, where it is a pre-requisite for such systems.



# Appendix **A**

## Ethical Approval

During this study, 7 pigs (*Sus scrofa domesticus*; ssp Large White), were involved in this non-survival experimental study. The protocol received full approval from the local Ethical Committee for animal use and care (ICOMETH; protocol n 38.2015.01.069, acronym ETICA) and was approved by the French Ministry of Superior Education and Research (MESR) under the reference number 2015092210412678 v4 APAFIS#1830). Animals were managed according to the ARRIVE guidelines (Kilkenny et al., 2010) and in accordance with French laws for animal use and care, and according to the directives of the European Community Council (2010/63/EU). Before the procedure, animals were placed in individual cages with controlled conditions of light/dark cycles, humidity and temperature as per regulatory standards. A standardized nutrition was provided and cages were enriched with toys. Pigs were fasted for 24 hours before surgery with free access to water. Ten minutes before surgery, animals were premedicated with an intramuscular injection of ketamine (20mg/Kg) and azaperone (2mg/Kg) (Stresnil, Janssen-Cilag, Belgium). Induction was achieved using intravenous propofol (3mg/Kg) combined with rocuronium (0.8mg/Kg). Anaesthesia was maintained with 2% isoflurane. At the end of the experimental procedures, the animals were sacrificed with an intravenous injection of a lethal dose of potassium chloride.



# Bibliography

- Aanæs, H., Fisker, R., Åström, K., and Carstensen, J. M. (2002). “Robust factorization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (9), 1215–1225.
- Agarwal, S., M., K., and Others. *Ceres Solver*. <http://ceres-solver.org>.
- Agarwal, S., Snavely, N., Seitz, S. M., and Szeliski, R. (2010). “Bundle Adjustment in the Large”. In: *European Conference on Computer Vision: Part II*, pp. 29–42.
- Ahmed, A. H. and Farag, A. A. (2007). “Shape from Shading Under Various Imaging Conditions”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D., and Silva, C. T. (2003). “Computing and rendering point set surfaces”. In: *IEEE Transactions on Visualization and Computer Graphics*, vol. 9 (1), pp. 3–15.
- Amir-Khalili, A., Peyrat, J.-M., Hamarneh, G., and Abugharbieh, R. (2013). “3D Surface Reconstruction of Organs Using Patient-Specific Shape Priors in Robot-Assisted Laparoscopic Surgery”. In: *Abdominal Imaging: Computational and Clinical Applications*, pp. 184–193.
- Atasoy, S., Noonan, D. P., Benhimane, S., Navab, N., and Yang, G.-Z. (2008). “A Global Approach for Automatic Fibroscopic Video Mosaicing in Minimally Invasive Diagnosis”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 850–857.
- Aujol, J.-F. (2009). “Some First-Order Algorithms for Total Variation Based Image Restoration”. In: *Journal of Mathematical Imaging and Vision*, vol. 34 (3), pp. 307–327.
- Banks, J., Bennamoun, M., and Corke, P. (1997). “Non-parametric techniques for fast and robust stereo matching”. In: *IEEE Conference on Speech and Image Technologies for Computing and Telecommunications*, vol. 1, pp. 365–368.
- Bano, J., Nicolau, S.-A., Hostettler, A., Doignon, C., Marescaux, J., and Soler, L. (2013). “Registration of Preoperative Liver Model for Laparoscopic Surgery from Intraoperative 3D Acquisition”. In: *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pp. 201–210.
- Bartoli, A., Gerard, Y., Chadebecq, F., Collins, T., and Pizarro, D. (2015). “Shape-from-Template”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37 (10), pp. 2099–2118.
- Bendet, S. *Laparoscopic stomach surgery*. [Online; accessed 29-March-2018. marked as public domain, more details on Wikimedia Commons]. URL: [https://commons.wikimedia.org/wiki/File:Laparoscopic\\_stomach\\_surgery.jpg](https://commons.wikimedia.org/wiki/File:Laparoscopic_stomach_surgery.jpg).

- Bernhardt, S., Abi-Nahed, J., and Abugharbieh, R. (2013). “Robust Dense Endoscopic Stereo Reconstruction for Minimally Invasive Surgery”. In: *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pp. 254–262.
- Bernhardt, S., Nicolau, S. A., Soler, L., and Doignon, C. (2017). “The status of augmented reality in laparoscopic surgery as of 2016”. In: *Medical Image Analysis*, 37, pp. 66–90.
- Blaus, B. *Medical gallery of Blausen Medical 2014*. [Online; accessed 29-March-2018. Wiki-Journal of Medicine 1 (2).] URL: [https://commons.wikimedia.org/wiki/File:Blausen\\_0602\\_Laparoscopy\\_02.png](https://commons.wikimedia.org/wiki/File:Blausen_0602_Laparoscopy_02.png).
- Bonarini, A., Burgard, W., Fontana, G., Matteucci, M., Sorrenti, D. G., and Tardos, J. D. (2006). “RAWSEEDS: Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets”. In: *In proceedings of IROS Workshop on Benchmarks in Robotics Research*. Vol. On line.
- Burschka, D., Li, M., Taylor, R., and Hager, G. D. (2005). “Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery”. In: *Medical Image Analysis*, 9, pp. 413–426.
- Bylow, E., Sturm, J., Kerl, C., Kahl, F., and Cremers, D. (2013). “Real-Time Camera Tracking and 3D Reconstruction Using Signed Distance Functions”. In: *Robotics: Science and Systems*, vol. 2.
- Cadena, C. et al. (2016). “Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age”. In: *IEEE Transactions on Robotics*, vol. 32 (6), pp. 1309–1332.
- Chambolle, A. and Pock, T. (2011). “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging”. In: *Journal of Mathematical Imaging and Vision*, vol. 40 (1), pp. 120–145.
- Chang, J. Y., Park, H., Park, I. K., Lee, K. M., and Lee, S. U. (2011). “GPU-friendly multi-view stereo reconstruction using surfel representation and graph cuts”. In: *Computer Vision and Image Understanding*, vol. 115 (5), pp. 620–634.
- Chang, P.-L., Chen, D., Cohen, D., and Edwards, P. E. (2012a). “2D/3D Registration of a Preoperative Model with Endoscopic Video Using Colour-Consistency”. In: *Augmented Environments for Computer-Assisted Interventions*, pp. 1–12.
- Chang, P.-L., Stoyanov, D., Davison, A. J., and Edwards, P. E. (2013). “Real-Time Dense Stereo Reconstruction Using Convex Optimisation with a Cost-Volume for Image-Guided Robotic Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 42–49.
- Chang, P.-L., Handa, A., Davison, A. J., Stoyanov, D., and Edwards, P. E. (2014). “Robust Real-Time Visual Odometry for Stereo Endoscopy Using Dense Quadrifocal Tracking”. In: *Information Processing in Computer-Assisted Interventions*, pp. 11–20.
- Chang, Y.-Z., Hou, J.-F., Tsao, Y. H., and Lee, S.-T. (2012b). “Application of real-time single camera SLAM technology for image-guided targeting in neurosurgery”. In: *SPIE, Applications of Digital Image Processing XXXV*, vol. 8499.
- Chen, L., T., W., and John, N. W. (2017). “Real-time Geometry-Aware Augmented Reality in Minimally Invasive Surgery”. In: *Healthcare Technology Letters*, vol. 4 (5), pp. 163–167.
- Chen, L., Tang, W., John, N. W., Wan, T. R., and Zhang, J. J. (2018). “SLAM-based dense surface reconstruction in monocular Minimally Invasive Surgery and its application to Augmented Reality”. In: *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 135–146.

- Chen, X. et al. (2015). “Development of a surgical navigation system based on augmented reality using an optical see-through head-mounted display”. In: *Journal of Biomedical Informatics*, vol. 55, pp. 124–131.
- Chiuso, A., Favaro, P., Jin, H., and Soatto, S. (2000). “MFm: 3-D Motion From 2-D Motion Causally Integrated Over Time”. In: *European Conference on Computer Vision*, pp. 735–750.
- Civera, J., Davison, A. J., and Montiel, J. M. M. (2008). “Inverse Depth Parametrization for Monocular SLAM”. In: *IEEE Transactions on Robotics*, vol. 24 (5), pp. 932–945.
- Civera, J., Grasa, Ó. G., Davison, A. J., and Montiel, J. M. M. (2009). “1-point RANSAC for EKF-based Structure from Motion”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3498–3504.
- Collins, R. T. (1996). “A space-sweep approach to true multi-image matching”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 358–363.
- Collins, T. and Bartoli, A. (2012a). “3D Reconstruction in Laparoscopy with Close-range Photometric Stereo”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 634–642.
- (2012b). “Towards Live Monocular 3D Laparoscopy Using Shading and Specularity Information”. In: *Information Processing in Computer-Assisted Interventions*, pp. 11–21.
- Collins, T., Pizarro, D., Bartoli, A., Canis, M., and Bourdel, N. (2013). “Realtime Wide-Baseline Registration of the Uterus in Laparoscopic Videos Using Multiple Texture Maps”. In: *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pp. 162–171.
- Concha, A. and Civera, J. (2015). “DPPTAM: Dense piecewise planar tracking and mapping from a monocular sequence”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5686–5693.
- Crandall, D., Owens, A., Snavely, N., and Huttenlocher, D. (2011). “Discrete-continuous optimization for large-scale structure from motion”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3001–3008.
- Davison, A. J. (2003). “Real-Time Simultaneous Localisation and Mapping with a Single Camera”. In: *IEEE International Conference on Computer Vision*, vol. 2.
- Deguchi, K. and Okatani, T. (1996). “Shape reconstruction from an endoscope image by shape-from-shading technique for a point light source at the projection center”. In: *Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 290–298.
- Engel, J., Schöps, T., and Cremers, D. (2014). “LSD-SLAM: Large-Scale Direct Monocular SLAM”. In: *European Conference on Computer Vision*, pp. 834–849.
- Engel, J., Koltun, V., and Cremers, D. (2018). “Direct Sparse Odometry”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40 (3), pp. 611–625.
- Esteban, C. H., Vogiatzis, G., and Cipolla, R. (2008). “Multiview Photometric Stereo”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30 (3), pp. 548–554.
- Forster, C. H., Quartucci and Tozzi, C. L. (2000). “Towards 3D reconstruction of endoscope images using shape from shading”. In: *Proceedings Brazilian Symposium on Computer Graphics and Image Processing*, pp. 90–96.
- Furukawa, Y. and Hernández, C. (2015). “Multi-View Stereo: A Tutorial”. In: *Foundations and Trends in Computer Graphics and Vision*, vol. 9 (1-2), pp. 1–148.

- Gallup, D., Frahm, J. M., Mordohai, P., Yang, Q., and Pollefeys, M. (2007). “Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Galvez-López, D. and Tardos, J. D. (2012). “Bags of Binary Words for Fast Place Recognition in Image Sequences”. In: *IEEE Transactions on Robotics*, vol. 28 (5), pp. 1188–1197.
- Gao, X.-S., Hou, X.-R., Tang, J., and Cheng, H.-F. (2003). “Complete solution classification for the perspective-three-point problem”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25 (8), pp. 930–943.
- Graber, G., Pock, T., and Bischof, H. (2011). “Online 3D reconstruction using convex optimization”. In: *IEEE International Conference on Computer Vision Workshops*, pp. 708–711.
- Grasa, Ó. G., Civera, J., G., A., and V. M., J.M.M. Montiel (2009). “EKF Monocular SLAM 3D Modeling, Measuring and Augmented Reality from Endoscope Image Sequences”. In: *Workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery*.
- Grasa, Ó. G., Civera, J., and Montiel, J. M. M. (2011). “EKF monocular SLAM with relocalization for laparoscopic sequences”. In: *IEEE International Conference on Robotics and Automation*, pp. 4816–4821.
- Grasa, Ó. G., Bernal, E., Casado, S., Gil, I., and Montiel, J. M. M. (2014). “Visual SLAM for Handheld Monocular Endoscope”. In: *IEEE Transactions on Medical Imaging*, vol. 33 (1), pp. 135–146.
- Hallet, J. et al. (2015). “Trans-Thoracic Minimally Invasive Liver Resection Guided by Augmented Reality”. In: *Journal of the American College of Surgeons*, vol. 220 (5), pp. 55–60.
- Han, M. and Kanade, T. (2003). “Multiple motion scene reconstruction with uncalibrated cameras”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (7), pp. 884–894.
- Handa, A., Newcombe, R. A., Angeli, A., Davison, A. J., and London, S. A. (2011). *Applications of Legendre-Fenchel transformation to computer vision problems*. Technical Report DTR11-7. London:Imperial College.
- Harris, C. G. and Pike, J. M. (1988). “3D Positional Integration from Image Sequences”. In: *Image and Vision Computing*, vol. 6 (2), pp. 87–90.
- Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. 2nd ed. New York, NY, USA: Cambridge University Press.
- Heise, P., Jensen, B., Klose, S., and Knoll, A. (2015). “Variational PatchMatch MultiView Reconstruction and Refinement”. In: *IEEE International Conference on Computer Vision*, pp. 882–890.
- Hirschmuller, H. (2008). “Stereo Processing by Semiglobal Matching and Mutual Information”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30 (2), pp. 328–341.
- Hirschmuller, H. and Scharstein, D. (2009). “Evaluation of Stereo Matching Costs on Images with Radiometric Differences”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31 (9), pp. 1582–1599.
- Horn, Berthold K. P. (1987). “Closed-form solution of absolute orientation using unit quaternions”. In: *Journal of the Optical Society of America A*, vol. 4 (4), pp. 629–642.

- Hu, M., Penney, G., Edwards, P., Figl, M., and Hawkes, D. J. (2007). “3D Reconstruction of Internal Organ Surfaces for Minimal Invasive Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 68–77.
- Hu, M. et al. (2012). “Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes”. In: *Medical Image Analysis*, vol. 16 (3), pp. 597–611.
- Karargyris, A. and Bourbakis, N. G. (2011). “Three-Dimensional Reconstruction of the Digestive Wall in CapsuleEndoscopy Videos Using Elastic Video Interpolation”. In: *IEEE Transaction on Medical Imaging*, vol. 30 (4), pp. 957–971.
- Kazhdan, M., Bolitho, M., and Hoppe, H. (2006). “Poisson Surface Reconstruction”. In: *Eurographics Symposium on Geometry Processing*, pp. 61–70.
- Kilgus, T. et al. (2015). “Mobile markerless augmented reality and its application in forensic medicine”. In: *International Journal of Computer Assisted Radiology and Surgery*, vol. 10 (5), pp. 573–586.
- Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). “Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research”. In: *PLOS Biology*, vol. 8 (6), pp. 1–5.
- Klein, G. and Murray, D. (2007). “Parallel Tracking and Mapping for Small AR Workspaces”. In: *IEEE and ACM International Symposium on Mixed and Augmented Reality*.
- Köhler, T. et al. (2013). “ToF Meets RGB: Novel Multi-Sensor Super-Resolution for Hybrid 3-D Endoscopy”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 139–146.
- Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., and Burgard, W. (2011). “G2o: A general framework for graph optimization”. In: *IEEE International Conference on Robotics and Automation*, pp. 3607–3613.
- Langguth, F., Sunkavalli, K., Hadap, S., and Goesele, M. (2016). “Shading-aware Multi-view Stereo”. In: *European Conference on Computer Vision*.
- Lapeer, R., Chen, M. S., Gonzalez, G., Linney, A., and Alusi, G. (2008). “Image-enhanced surgical navigation for endoscopic sinus surgery: evaluating calibration, registration and tracking”. In: *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 4 (1), pp. 32–45.
- Lee, J.-D., Huang, C.-H., Huang, T.-C., Hsieh, H.-Y., and Lee, S.-T. (2012). “Medical augment reality using a markerless registration framework”. In: *Expert Systems with Applications*, vol. 39 (5), pp. 5286–5294.
- Lepetit, V., Moreno-Noguer, F., and Fua, P. (2009). “EPnP: An Accurate  $O(n)$  Solution to the PnP Problem”. In: *International Journal Computer Vision*, vol. 81 (2).
- Lin, B., Johnson, A., Qian, X., Sanchez, J., and Sun, Y. (2013). “Simultaneous Tracking, 3D Reconstruction and Deforming Point Detection for Stereoscope Guided Surgery”. In: *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pp. 35–44.
- Lin, B., Sun, Y., Qian, X., Goldgof, D., Gitlin, R., and You, Y. (2016). “Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey”. In: *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 12 (2), pp. 158–178.

- Lin, J., Clancy, N. T., and Elson, D. S. (2015). “An endoscopic structured light system using multispectral detection”. In: *International Journal of Computer Assisted Radiology and Surgery*, vol. 10 (12), pp. 1941–1950.
- Longuet-Higgins, H. C. (1981). “A computer algorithm for reconstructing a scene from two projections,” in: *Nature*, vol.293 (12), pp. 133–135.
- Lowe, David G. (2004). “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision*, vol. 60 (2), pp. 91–110.
- Lucas, B. D. and Kanade, T. (1981). “An Iterative Image Registration Technique with an Application to Stereo Vision”. In: *International Joint Conference on Artificial Intelligence*, vol. 2, pp. 674–679.
- Macedo, M. C. F., Júnior, A. L. A., Santos Souza, A. C. dos, and Giralaldi, G. A. (2014). “High-quality on-patient medical data visualization in a markerless augmented reality environment”. In: *SBC Journal on Interactive Systems*, vol. 5 (3), pp. 1101 –1114.
- Mahmoud, N., Nicolau, S. A., Keshk, A., Ahmad, M. A., Soler, L., and Marescaux, J. (2012). “Fast 3D Structure From Motion with Missing Points from Registration of Partial Reconstructions”. In: *Articulated Motion and Deformable Objects*, pp. 173–183.
- Mahmoud, N., Grasa, Ó. G., Nicolau, S. A., Doignon, C., Soler, L., Marescaux, J., and Montiel, J. M. M. (2017a). “On-patient see-through augmented reality based on visual SLAM”. In: *International Journal of Computer Assisted Radiology and Surgery*, vol. 12 (1), pp. 1–11.
- Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., and Montiel, J. M. M. (2017b). “ORB-SLAM-Based Endoscope Tracking and 3D Reconstruction”. In: *MICCAI Workshop Computer-Assisted and Robotic Endoscopy*. Springer International Publishing, pp. 72–83.
- Mahmoud, N., Hostettler, A., Collins, T., Doignon, C., Soler, L., and Montiel, J. M. M. (2017c). “SLAM based Quasi Dense Reconstruction For Minimally Invasive Surgery Scenes”. In: *ICRA 2017 workshop C4 Surgical Robots: Compliant, Continuum, Cognitive, and Collaborative*. *arXiv:1705.09107*.
- Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., and Montiel, J. M. M. (2018). “Live Tracking and Dense Reconstruction for Hand-held Monocular Endoscopy”. In: *IEEE Transation on Medical Imaging (Submitted)*.
- Mahmoud, Na., Collins, T., Hostettler, A., Soler, L., Doignon, C., and Montiel, J. M. M. (2017d). “Quasi-Dense Reconstruction from Monocular Laparoscopic Video”. In: *Surgetica conference*.
- Maier-Hein, L. et al. (2014a). “Can Masses of Non-Experts Train Highly Accurate Image Classifiers?” In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 438–445. URL: <http://opencas.webarchiv.kit.edu/?q=InstrumentCrowd>.
- Maier-Hein, L. et al. (2014b). “Comparative Validation of Single-Shot Optical Techniques for Laparoscopic 3-D Surface Reconstruction”. In: *IEEE Transactions on Medical Imaging*, vol. 33 (10), pp. 1913–1930. URL: <http://opencas.webarchiv.kit.edu/?q=tmidataset>.
- Malti, A., Bartoli, A., and Collins, T. (2012). “Template-Based Conformal Shape-from-Motion-and-Shading for Laparoscopy”. In: *Information Processing in Computer-Assisted Interventions*, pp. 1–10. URL: <http://igt.ip.uca.fr/~ab/Research/index.html>.
- Marcinczak, J. M. and Grigat, R.-R. (2014). “Total Variation Based 3D Reconstruction from Monocular Laparoscopic Sequences”. In: *Abdominal Imaging: Computational and Clinical Applications*, pp. 239–247.

- Martinez-Herrera, S. E., Malti, A., Morel, O., and Bartoli, A. (2013). “Shape-from-Polarization in laparoscopy”. In: *IEEE International Symposium on Biomedical Imaging*, pp. 1412–1415.
- Maurice, X., Albitar, C., Doignon, C., and Mathelin, M. de (2012). “A structured light-based laparoscope with real-time organs’ surface reconstruction for minimally invasive surgery”. In: *IEEE Engineering in Medicine and Biology Conference*.
- McCormac, J., Handa, A., Davison, A. J., and Leutenegger, S. (2017). “SemanticFusion: Dense 3D semantic mapping with convolutional neural networks”. In: *IEEE International Conference on Robotics and Automation*, pp. 4628–4635.
- Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., and Bartoli, A. (2016). “Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy”. In: *IEEE Transactions on Medical Imaging*, vol. 35 (9), pp. 2051–2063.
- Mileva, Y., Bruhn, A., and Weickert, J. (2007). “Illumination-Robust Variational Optical Flow with Photometric Invariants”. In: *Pattern Recognition. DAGM 2007*, vol. 4713, pp. 152–162.
- Mirota, D., Wang, H., Taylor, R. H., Ishii, M., and Hager, G. D. (2009). “Toward Video-Based Navigation for Endoscopic Endonasal Skull Base Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 91–99.
- Mirota, D. J., Wang, H., Taylor, R. H., Ishii, M., Gallia, G. L., and Hager, G. D. (2012). “A System for Video-Based Navigation for Endoscopic Endonasal Skull Base Surgery”. In: *IEEE Transactions on Medical Imaging*, vol. 31 (4), pp. 963–976.
- Montiel, J. M. M., Civera, J., and Davison, A. J. (2006). “Unified inverse depth parametrization for monocular SLAM”. In: *Proceedings of Robotics: Science and Systems*.
- Mountney, P., Stoyanov, D., Davison, A., and Yang, G.-Z. (2006). “Simultaneous Stereoscope Localization and Soft-Tissue Mapping for Minimal Invasive Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 347–354.
- Mountney, P. and Yang, G. Z. (2009). “Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping”. In: *IEEE International Conference Engineering in Medicine and Biology Society*, pp. 1184–1187.
- (2010a). “Motion Compensated SLAM for Image Guided Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 496–504.
- Mountney, P., Stoyanov, D., and Yang, G. Z. (2010b). “Three-Dimensional Tissue Deformation Recovery and Tracking,” in: *IEEE Signal Processing Magazine* vol. 27 (4), pp. 14–24. URL: <http://hamlyn.doc.ic.ac.uk/vision/>.
- Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., and Sayd, P. (2006). “Real Time Localization and 3D Reconstruction”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 363–370.
- Müller, M. et al. (2013). “Mobile augmented reality for computer-assisted percutaneous nephrolithotomy”. In: *International Journal of Computer Assisted Radiology and Surgery*, vol. 8 (4), pp. 663–675.
- Mur-Artal, R., Montiel, J. M. M., and Tardós, J. D. (2015). “ORB-SLAM: A Versatile and Accurate Monocular SLAM System”. In: *IEEE Transactions on Robotics*, vol. 31 (5), pp. 1147–1163.
- Mur-Artal, R. and Tardós, J. D. (2015). “Probabilistic Semi-Dense Mapping from Highly Accurate Feature-Based Monocular SLAM”. In: *Robotics: Science and Systems XI, Sapienza University of Rome*.

- Navab, N., Traub, J., Sielhorst, T., Feuerstein, M., and Bichlmeier, C. (2007). “Action- and Workflow-Driven Augmented Reality for Computer-Aided Medical Procedures”. In: *IEEE Computer Graphics and Applications*, vol. 27 (5), pp. 10–14.
- Neira, J. and Tardos, J. D. (2001). “Data association in stochastic mapping using the joint compatibility test”. In: *IEEE Transactions on Robotics and Automation*, vol. 17 (6), pp. 890–897.
- Newcombe, R. A. and Davison, A. J. (2010). “Live dense reconstruction with a single moving camera”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1498–1505.
- Newcombe, R. A., Lovegrove, S. J., and Davison, A. J. (2011a). “DTAM: Dense tracking and mapping in real-time”. In: *International Conference on Computer Vision*, pp. 2320–2327.
- Newcombe, R. A. et al. (2011b). “KinectFusion: Real-time dense surface mapping and tracking”. In: *IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136.
- Nicolau, S.A., Pennec, X., Soler, L., Buy, X., Gangi, A., Ayache, N., and Marescaux, J. (2009). “An augmented reality system for liver thermal ablation: Design and evaluation on clinical cases”. In: *Medical Image Analysis*, vol. 13 (3), pp. 494–506.
- Noonan, D. P., Mountney, P., Elson, D. S., Darzi, A., and Yang, G. Z. (2009). “A stereoscopic fibroscope for camera motion and 3D depth recovery during Minimally Invasive Surgery”. In: *IEEE International Conference on Robotics and Automation*, pp. 4463–4468.
- Opfermann, J. D., Leonard, S., Decker, R. S., Uebele, N. A., Bayne, C. E., Joshi, A. S., and Krieger, A. (2017). “Semi-autonomous electrosurgery for tumor resection using a multi-degree of freedom electrosurgical tool and visual servoing”. In: *International Conference on Intelligent Robots and Systems*, pp. 3653–3660.
- Penza, V., Ortiz, J., Mattos, L. S., Forgione, A., and De Momi, E. (2016). “Dense soft tissue 3D reconstruction refined with super-pixel segmentation for robotic abdominal surgery”. In: *International Journal of Computer Assisted Radiology and Surgery*, vol. 11 (2), pp. 197–206.
- Pizzoli, M., Forster, C., and Scaramuzza, D. (2014). “REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time”. In: *IEEE International Conference on Robotics and Automation*.
- Poelman, C. J. and Kanade, T. (1997). “A paraperspective factorization method for shape and motion recovery”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19 (3), pp. 206–218.
- Prados, E., Camilli, F., and Faugeras, O. (2006). “A Unifying and Rigorous Shape from Shading Method Adapted to Realistic Data and Applications”. In: *Journal of Mathematical Imaging and Vision*, vol. 25 (3), pp. 307–328.
- Pratt, P., Bergeles, C., Darzi, A., and Yang, G.-Z. (2014). “Practical Intraoperative Stereo Camera Calibration”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 667–675.
- Puerto-Souza, G. A. and Mariottini, G. L. (2013). “A Fast and Accurate Feature-Matching Algorithm for Minimally-Invasive Endoscopic Images”. In: *IEEE Transactions on Medical Imaging*, vol. 32 (7), pp. 1201–1214. URL: [http://ranger.uta.edu/~gianluca/feature\\_matching/](http://ranger.uta.edu/~gianluca/feature_matching/).
- Quan, L. and Kanade, T. (1996). “A factorization method for affine structure from line correspondences”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 803–808.

- Queàu, Y., Mérou, J., Durou, J.-D., and Cremers, D. (2017). “Dense Multi-view 3D-reconstruction Without Dense Correspondences”. In: *ArXiv preprint 1704.00337*.
- Rassweiler, J. J. et al. (2012). “iPad-Assisted Percutaneous Access to the Kidney Using Marker-Based Navigation: Initial Clinical Experience”. In: *European Urology*, vol. 61 (3), pp. 628–631.
- Rosten, E. and Drummond, T. (2006). “Machine Learning for High-Speed Corner Detection”. In: *European Conference on Computer Vision*, pp. 430–443.
- Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. R. (2011). “ORB: An efficient alternative to SIFT or SURF”. In: *IEEE International Conference on Computer Vision*, pp. 2564–2571.
- Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008). “Towards 3D Point cloud based object maps for household environments”. In: *Robotics and Autonomous Systems*, vol. 56 (11), pp. 927–941.
- Röhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Müller-Stich, B. P., Dillmann, R., and Speidel, S. (2012). “Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration”. In: *Medical Physics*, vol. 39 (3), pp. 1632–1645. URL: <http://opencas.webarchiv.kit.edu/?q=data/simulationdata>.
- S., Agarwal, Y., Furukawa, N., Snavely, I., Simon, B., Curless, M., Seitz S., and R., Szeliski (2011). “Building Rome in a Day”. In: *Communication of the ACM*. vol. 54 (10), pp. 105–112.
- Santos, T. R. dos et al. (2014). “Pose-independent surface matching for intra-operative soft-tissue marker-less registration”. In: *Medical Image Analysis*, vol. 18 (7), pp. 1101–1114.
- Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesci, N., Wang, X., and Westling, P. (2014). “High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth.” In: *German Conference on Pattern Recognition*. Vol. 8753, pp. 31–42. URL: <http://vision.middlebury.edu/stereo/data/scenes2014/>.
- Schneider, A., Baumberger, C., Griessen, M., Pezold, S., Beinemann, J., Jürgens, P., and Cattin, P. C. (2014). “Landmark-Based Surgical Navigation”. In: *Clinical Image-Based Procedures. Translational Research in Medical Imaging*, pp. 57–64.
- Schoob, A., Podzus, F., Kundrat, D., Kahrs, L. A., and Ortmaier, T. (2013). “Stereoscopic Surface Reconstruction in Minimally Invasive Surgery using Efficient Non-Parametric Image Transforms”. In: *Workshop on New Technologies for Computer/Robot Assisted Surgery*.
- Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). “A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 519–528.
- Shi, J. and Tomasi, C. (1994). “Good features to track”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600.
- Shum, H.-Y., Ke, Q., and Zhang, Z. (1999). “Efficient bundle adjustment with virtual key frames: a hierarchical approach to multi-frame structure from motion”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 543.
- Sielhorst, T., Feuerstein, M., Traub, J., Kutter, O., and Navab, N. (2006). “CAMPAR: A software framework guaranteeing quality for medical augmented reality”. In: *International Journal of Computer Assisted Radiology and Surgery*, pp. 29–30.
- Snavely, N., Seitz, S. M., and Szeliski, R. (2006). “Photo Tourism: Exploring Photo Collections in 3D”. In: *ACM Transactions on Graphics*, vol. 25 (3), pp. 835–846.

- Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G. (2018). “Dynamic Reconstruction of Deformable Soft-Tissue With Stereo Scope in Minimal Invasive Surgery”. In: *IEEE Robotics and Automation Letters*, vol. 3 (1), pp. 155–162.
- Speidel, S., Kenngott, H., and Maier-Hein, L. In: *Open-CAS: validating and benchmarking computer-assisted surgery systems*. URL: <http://opencas.webarchiv.kit.edu/>.
- Stoyanov, D., Darzi, A., and Yang, G.-Z. (2005). “Laparoscope Self-calibration for Robotic Assisted Minimally Invasive Surgery”. In: *Medical Image Computing and Computer Assisted*, pp. 114–121.
- Stoyanov, D., Scarzanella, M. V., Pratt, P., and Yang, G.-Z. (2010). “Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 275–282. URL: <http://hamlyn.doc.ic.ac.uk/vision/>.
- Stoyanov, Danail (2012). “Surgical Vision”. In: *Annals of Biomedical Engineering*, vol. 40 (2), pp. 332–345.
- Strasdat, H., Montiel, J. M. M., and Davison, A. J. (2010). “Scale Drift-Aware Large Scale Monocular SLAM”. In: *Robotics: Science and Systems VI, Universidad de Zaragoza, Zaragoza, Spain*.
- Strasdat, H., Davison, A. J., Montiel, J. M. M., and Konolige, K. (2011). “Double window optimisation for constant time visual SLAM”. In: *International Conference on Computer Vision*, pp. 2352–2359.
- Strasdat, H., Montiel, J. M. M., and Davison, A. J. (2012). “Visual SLAM: Why Filter?” In: *Image and Vision Computing*, vol. 30 (2), pp. 65–77.
- Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. (2012). “A Benchmark for the Evaluation of RGB-D SLAM Systems”. In: *International Conference on Intelligent Robot Systems*. URL: <https://vision.in.tum.de/data/datasets/rgbd-dataset/download>.
- Su, L.-M., Vagvolgyi, B.-P., Agarwal, R., Reiley, C.-E., Taylor, R.-H., and Hager, G.-D. (2009). “Augmented Reality During Robot-assisted Laparoscopic Partial Nephrectomy: Toward Real-Time 3D-CT to Stereoscopic Video Registration”. In: *Urology*, vol. 73 (4), pp. 896–900.
- Sugimoto, M. et al. (2010). “Image overlay navigation by markerless surface registration in gastrointestinal, hepatobiliary and pancreatic surgery”. In: *Journal of Hepato-Biliary-Pancreatic Sciences*, vol. 17 (5), pp. 629–636.
- Sun, D., Liu, J., Linte, C. A., Duan, H., and Robb, R. A. (2013). “Surface Reconstruction from Tracked Endoscopic Video Using the Structure from Motion Approach”. In: *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pp. 127–135.
- Surgical Robot Vision*. <http://www.surgicalvision.cs.ucl.ac.uk/resources/benchmarking/#tracking>. Accessed: 26-03-2018.
- Suwelack, S. et al. (2014). “Physics-based shape matching for intraoperative image guidance.” In: *Medical physics*, 41. URL: <http://opencas.webarchiv.kit.edu/?q=PhysicsBasedShapeMatching>
- Szeliski, Richard (2010). *Computer Vision: Algorithms and Applications*. 1st. Springer-Verlag New York, Inc.
- Tankus, A., Sochen, N., and Yeshurun, Y. (2004). “Reconstruction of medical images by perspective shape-from-shading”. In: *International Conference on Pattern Recognition*, vol. 3, pp. 778–781.

- (2005). “Shape-from-Shading Under Perspective Projection”. In: *International Journal of Computer Vision*, vol. 63 (1), pp. 21–43.
- Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., and Pollefeys, M. (2013). “Live Metric 3D Reconstruction on Mobile Phones”. In: *IEEE International Conference on Computer Vision*.
- Tateno, K., Tombari, F., Laina, I., and Navab, N. (2017). “CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction”. In: *Computer Vision and Pattern Recognition*, pp. 404–412.
- Tomasi, C. and Kanade, T. (1992). “Shape and motion from image streams under orthography: a factorization method”. In: *International Journal of Computer Vision*, vol. 9 (2), pp. 137–154.
- Totz, J., Mountney, P., Stoyanov, D., and Yang, G.-Z. (2011). “Dense Surface Reconstruction for Enhanced Navigation in MIS”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 89–96.
- Totz, J., Thompson, S., Stoyanov, D., Gurusamy, K., Davidson, B. R., Hawkes, D. J., and Clarkson, M. J. (2014). “Fast Semi-dense Surface Reconstruction from Stereoscopic Video in Laparoscopic Surgery”. In: *Information Processing in Computer-Assisted Interventions*, pp. 206–215.
- Triggs, B. (1996). “Factorization methods for projective structure and motion”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 845–851.
- Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (2000). “Bundle Adjustment - A Modern Synthesis”. In: *International Workshop on Vision Algorithms: Theory and Practice*, pp. 298–372.
- Tromberg, B. J. et al. (2000). “Non-Invasive In Vivo Characterization of Breast Tumors Using Photon Migration Spectroscopy”. In: *Neoplasia*, vol. 2 (1), pp. 26–40.
- Tsai, P. s. and Shah, M. (1994). “Shape from shading using linear approximation”. In: *Image and Vision Computing*, pp. 487–498.
- Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., and Sitti, M. (2017). “A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots”. In: *International Journal of Intelligent Robotics and Applications*, vol. 1 (4), pp. 399–409.
- Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M. de, and Padoy, N. (2017). “EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos”. In: *IEEE Transaction on Medical Imaging*, vol. 36 (1), pp. 86–97.
- Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M. de, and Padoy, N. (2016). “EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos”. In: *IEEE Transactions on Medical Imaging*. URL: <http://camma.u-strasbg.fr/datasets>. Visible Patient. <https://www.visiblepatient.com/en/>. Accessed: 27-02-2018.
- Wang, H., Mirota, D., Ishii, M., and Hager, G. D. (2008). “Robust motion estimation and structure recovery from endoscopic image sequences with an Adaptive Scale Kernel Consensus estimator”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7. WebSurg. <https://www.websurg.com/>. Accessed: 28-02-2018.
- Wendel, A., Maurer, M., Graber, G., Pock, T., and Bischof, H. (2012). “Dense reconstruction on-the-fly”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1450–1457.

- Wu, C., Narasimhan, S. G., and Jaramaz, B. (2010). “A Multi-Image Shape-from-Shading Framework for Near-Lighting Perspective Endoscopes”. In: *International Journal of Computer Vision*, vol. 86 (2), pp. 211–228.
- Wu, C. H., Sun, Y. N., and Chang, C. C. (2007). “Three-Dimensional Modeling From Endoscopic Video Using Geometric Constraints Via Feature Positioning”. In: *IEEE Transactions on Biomedical Engineering*, vol. 54 (7), pp. 1199–1211.
- Wunderling, T., Golla, B., Poudel, P., Arens, C., Friebe, M., and Hansen, C. (2017). “Comparison of thyroid segmentation techniques for 3D ultrasound”. In: *Proceedings SPIE Medical Imaging*, vol. 10133, pp. 10133 –10133 –7.
- Yip, M. C., Lowe, D. G., Salcudean, S. E., Rohling, R. N., and Ngan, C. Y. (2012). “Tissue Tracking and Registration for Image-Guided Surgery”. In: *IEEE Transactions on Medical Imaging*, vol. 31 (11), pp. 2169–2182.
- Zhang, Z. (2000). “A flexible new technique for camera calibration”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22 (11), pp. 1330–1334.
- Zhou, Z., Wu, Z., and Tan, P. (2013). “Multi-view Photometric Stereo with Spatially Varying Isotropic Materials”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1482–1489.

# SLAM visual monocular para cirugía mínimamente invasiva con aplicación a realidad aumentada

Visual Monocular SLAM for Minimally Invasive Surgery and its application to  
Augmented Reality

Nader Mahmoud Elshahat Elsayed ALI<sup>1,3</sup>,  
**Director:** Prof. José María Martínez MONTIEL<sup>2</sup>, and  
Prof. Christophe DOIGNON<sup>3</sup>

<sup>1</sup>*IRCAD (Institut de Recherche contre les Cancers de l'Appareil Digestif), France.*

<sup>2</sup>*Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain.*

<sup>3</sup>*ICube (UMR 7357 CNRS), Université de Strasbourg, France.*

May 2018

## Resumen y Conclusiones

La cirugía mínimamente invasiva (MIS) ha ganado mucha popularidad en los últimos años. Durante el procedimiento quirúrgico se inserta una cámara endoscópica en la cavidad abdominal a través de pequeñas incisiones hechas en la piel del paciente, lo que permite mostrar sus estructuras internas en un monitor en el sala de operaciones. Los beneficios de los procedimientos MIS sobre la cirugía abierta tradicional son muchos: una estancia hospitalaria más corta, cicatrices más pequeñas (porque solo incisiones de unos pocos milímetros son necesarias), menos sangrado y, como resultado, un riesgo postquirúrgico es mucho menor. Sin embargo, aunque muchos beneficios del MIS son innegable para el paciente, el procedimiento es más difícil para el personal quirúrgico y requiere un mayor tiempo de aprendizaje. Las principales dificultades son las siguientes: 1) coordinación ojo-mano en una escena 3D observada a través de una pantalla 2D lo que implica un punto de vista diferente, por lo que es necesario superar el reflejo natural para dirigir el ojos hacia la actividad de las manos. 2) La pérdida de la visión directa, de hecho, el cirujano solo puede ver las imágenes 2D en el monitor y esto pueden causar percepciones erróneas, especialmente la percepción de profundidad que es crucial para evaluar adecuadamente la relación espacial entre tejidos. 3) El campo de visión (FOV) proporcionado por la cámara endoscópica es apreciablemente más pequeño que el de la visión directa durante la cirugía abierta. 4) Sólo visión endoscópica que no proporciona información sobre estructuras anatómicas críticas como tumores y vasos ya que están ubicados debajo de la superficie de los órganos. Por lo tanto, los procedimientos MIS puede ser particularmente tediosos y aumentan significativamente la duración de las operaciones. Para superar las limitaciones inherentes de las intervenciones de MIS, esta tesis se centra en en el estudio de la localización y mapeo denso a partir de del flujo de video de un endoscopio monocular estándar (en inglés SLAM - Simultaneous Localization And Mapping). Precisamente, en el contexto de una intervención MIS, los dos pasos básicos: 1) reconstrucción densa en 3D del campo quirúrgico y 2) la ubicación del endoscopio respecto de los órganos del paciente, deben realizarse en tiempo real. Los procedimientos de SLAM, desarrollados en el ámbito de la robótica y la visión por computadora, permiten una reconstrucción en 3D, denominada mapa, de una escena desconocida mientras que al mismo tiempo realiza un seguimiento de la trayectoria de la cámara en relación con el mapa. Esta información sobre ubicación del endoscopio y el mapeo es esencial para la computerización de los procedimientos MIS. La representación densa escena quirúrgica 3D compensa la limitación de campo de visión y mejora significativamente la percepción de la profundidad que puede tener el cirujano durante la intervención. Si se consigue que sea en tiempo, tenemos todos los ingredientes para diseñar un sistema completo de realidad aumentada (AR).

Mediante AR se busca superponer sobre la imagen endoscópica del paciente en el modelo geométrico preoperatorio del paciente. Las anotaciones de AR se crean a partir de diferentes modalidades de imágenes como tomografía computarizada (TC) o resonancia magnética marca (MRI). De esta manera, el cirujano tiene la sensación de ver a su paciente como si fuera transparente. Las anotaciones de AR puede proporcionar información decisiva durante el MIS, para la identificación estructuras intraoperatorias no visibles tales como tumores, vasos o nervios. Esto ayudaría al cirujano, que ya no tiene que adaptar mentalmente la información extraída de las imágenes médicas, para guiarse durante las resecciones mostrando las trayectorias y márgenes planificados previamente en modelos preoperatorios, o en la colocación óptima de trócares, también planificados preoperatoriamente.

El trabajo preliminar que presentamos en [1], introdujo un enfoque basado en SLAM, que permite una visualización intraoperatoria precisa de los modelos preoperatorios a través de la piel del paciente, usando solo un Tablet-PC. En este trabajo, una versión no densa de SLAM se utiliza para ubicar de manera robusta la cámara del Tablet-PC respecto del cuerpo del paciente. Para su empleo en la sala de operaciones, hemos desarrollado un conjunto de tratamientos, que van desde la grabación hasta la visualización, y que requieren interacciones mínimas del personal médico. De hecho, las interacciones con el cirujano se reducen a identificación de 4 a 6 referencias anatómicas al comienzo del procedimiento que se utilizan para registrar los datos preoperatorios. A diferencia de otros sistemas de AR, el sistema propuesto trackea la cámara, estima la reconstrucción de la escena observada, y renderiza la anotación de AR en tiempo real. Además, es robusto a una alta tasa de ocultaciones de la escena y no requiere dispositivos de seguimiento externos o marcadores artificiales en la piel del paciente. Este sistema ha sido rigurosamente evaluado en una serie de experimentos. Primero, el la precisión geométrica se evaluó mediante varios experimentos con marcadores añadidos tener el ground truth. Los primeros experimentos se llevaron a cabo in-vivo

en cerdos, los segundos en un maniquí. También se hizo una estimación del tiempo de cálculo con dos voluntarios, cada uno de ellos apoyado sobre una camilla mientras el personal médico sostenía el Tablet-PC y se movía alrededor. La Figura 1 muestra una visualización de AR, in vivo de datos preoperatorios del voluntario.

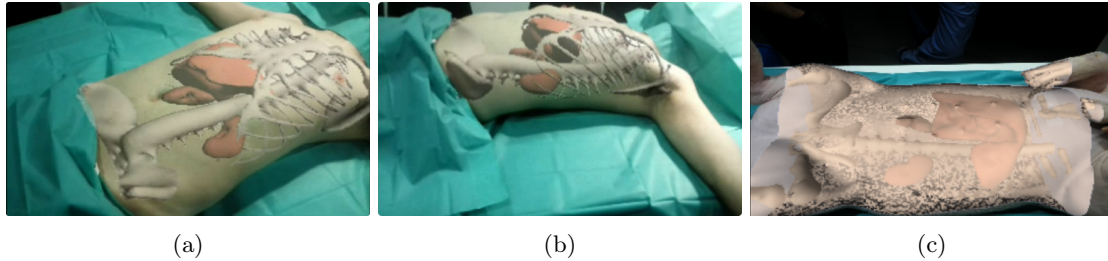


Figure 1: Visualización de realidad aumentada (AR) de datos preoperatorios, propuesta en [1]. (a-b) Superposición de los huesos, el hígado, los riñones izquierdo y derecho de un voluntario para dos puntos de vista. (c) Visualización con AR en un cerdo

En una segunda fase movemos el foco de interés hacia las secuencias de endoscopia médica para estimar un mapa denso de la escena y hacer tracking del endoscopio. Las dificultades a superar son las vastas áreas de la escenas que tienen una textura débil, y la cambiante iluminación de las superficies a lo largo de la secuencia, a veces incluyendo reflejos especulares. Además, la escena sufre deformaciones y el endoscopio sigue una trayectoria localmente abrupta dentro de la cavidad abdominal. Nuestra primera contribución a la explotación de SLAM visual en endoscopia es la configuración de los diversos parámetros del sistema popular ORB-SLAM para su funcionamiento en secuencias de endoscopia, el rendimiento de la parte de seguimiento es excelente, como lo demuestra el resultados reportados en [2]. Sin embargo, la calidad del mapa en términos de densidad es muy limitada. De hecho, la reconstrucción de la escena intracorporea tiene muy pocos puntos 3D debido al bajo número de puntos de interés repetibles (Fig. 2 (B)). La baja densidad del mapa resultante limita su uso para otras tareas diferentes de la localización del endoscopio en la cavidad abdominal. Nuestra segunda contribución consistió en proponer un algoritmo de reconstrucción que mejora significativamente la densidad del mapa y permite una reconstrucción robusta de la escena quirúrgica con una precisión de 4.9 mm [3, 4] (Fig. 2 (C)). Para ello, explotamos la fase de exploración inicial de la cavidad que típicamente es realizado por el cirujano al principio de la cirugía, de donde se seleccionan una conjunto de keyframes. Estos keyframes se usan para estimar la poses y el mapa mediante ajuste de haces, (Bundle Adjustment, BA). Después de la fase de exploración inicial, una vez localizados los keyframes, se asocian para formar pares de imágenes estéreo a partir de los que se estima un mapa denso de la escena. Gracias al seguimiento robusto del endoscopio y la representación de la escena densa obtenida pueden mostrarse las estructuras internas del hígado como anotaciones de realidad aumentada en tiempo real (Fig. 2 (d)). El único requisito es alinear los modelos preoperatorios con la reconstrucción densa (Fig. 2) (E)). Una vez hecho el alineamiento, la visualización de la AR se mantiene, mientras que la pose del endoscopio se calcula continuamente a medida que el endoscopio se mueve por la cavidad abdominal (Fig. 2 (f)).

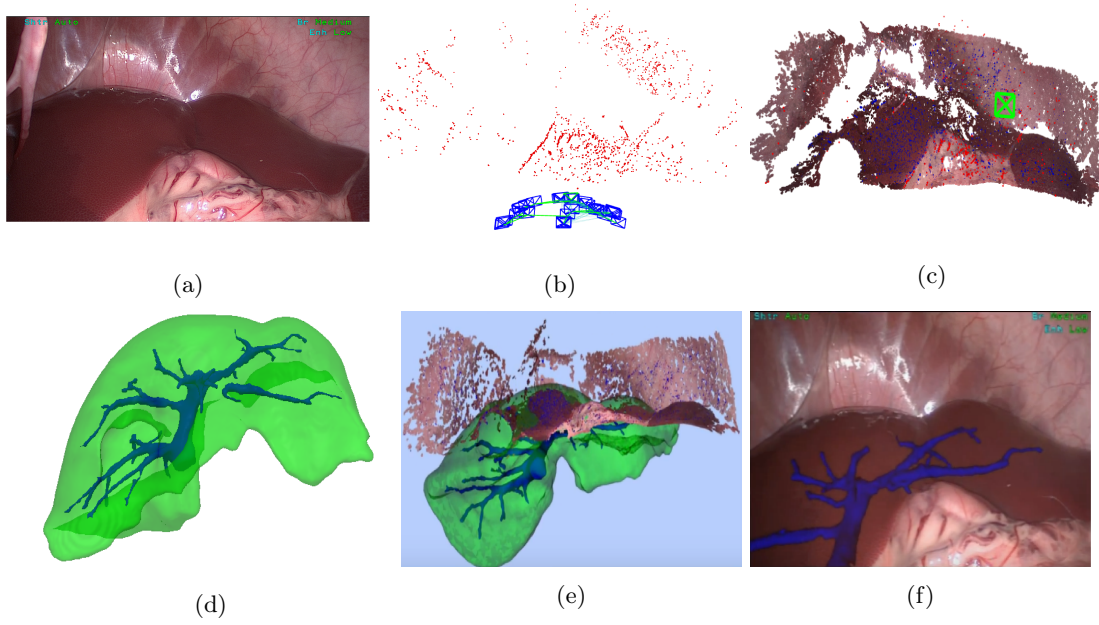


Figure 2: Reconstrucción de una escena intracorporea. (a) Una imagen típica de la secuencia. (b) Reconstrucción sparse mediante SLAM visual. (c) Reconstrucción densa. (d) Modelos de vena hepática segmentados semiautomáticamente a partir de imágenes de CT. (e) Registro modelo del preoperatorio del hígado con reconstrucción densa. (f) Visualización por RA sin marcadores de la vena hepática.

En [5], propusimos una solución para SLAM visual con reconstrucción densa en tiempo real, capaz de superar las dificultades propias de la endoscopia y que se ha aplicado con éxito en la laparoscopia in-vivo. El sistema propuesto amplía nuestro trabajo previo [?, 4] pero explotando múltiples vistas muy próximas para estimar la geometría densa del mapa. Por una parte, se ha hecho una implementación basada en multi-threading, añadiendo un nuevo hilo que realiza la reconstrucción densa de la escena densa, consiguiendo que se ejecute en vivo y en paralelo con otros hilos de tracking y mapping. Nuestra contribución consiste en que seleccionamos de todas las imágenes del video utilizado, un subconjunto de imágenes representativas de nuestra escena, llamado el conjunto keyframes. Luego, para cada uno de estos keyframes, se calcula un mapa denso de profundidad en usando imágenes contenidas en una pequeña ventana de tiempo justo antes y justo después de este keyframe. Estos mapas de profundidad se fusionan para crear una reconstrucción única. La ventana de tiempo se ajusta automáticamente para que el mapa de la profundidad en cuestión se pueda actualizar de forma robusta: normalmente, si la cámara se mueve lentamente (lo que dará como resultado un paralaje débil entre imágenes), la ventana se amplía para asegurar un buen condicionamiento geométrico. Para calcular el mapa de profundidad, usamos un método variacional que es robusto a fuertes cambios de iluminación por el empleo de la ZNCC (Zero Normalized Cross Correlation). Este enfoque contrasta con trabajos previos de SLAM denso que asumen invarianza en el brillo de cada píxel a lo largo del tiempo. Además, el método variacional que usamos permite reconstruir mapas de profundidad con discontinuidades, que se produce con frecuencia en escenas observadas por la laparoscopia donde los órganos y las herramientas quirúrgicas producen ocultaciones. El sistema completo ha sido validado experimentalmente y evaluado en secuencias de video de la cavidad abdominal de los cerdos. Además, su ejecución requiere un tiempo reducido gracias al uso de GPU. Comparado con otros métodos de SLAM visual denso obtenemos rendimientos superiores en términos de precisión y densidad. En conclusión, gracias a la selección eficiente de frames del video para generar una reconstrucción densa basada en criterios de paralaje, el método propuesto puede superar las prestaciones de las reconstrucciones binoculares, porque el grupo de frames puede proporcionar un paralaje mayor que el del endoscopio estereo. Figuras 3 (b, d) muestra los resultados de la reconstrucción densa e incremental de la superficie hepática, utilizada para calcular la visualización de AR (Fig. 3 (d)). La pose estimada del endoscopio (frustum rojo en la Fig. 3 (d)) se usa para actualizar el ángulo de vista del renderizado de la anotación de AR (Fig. 3 (e)).

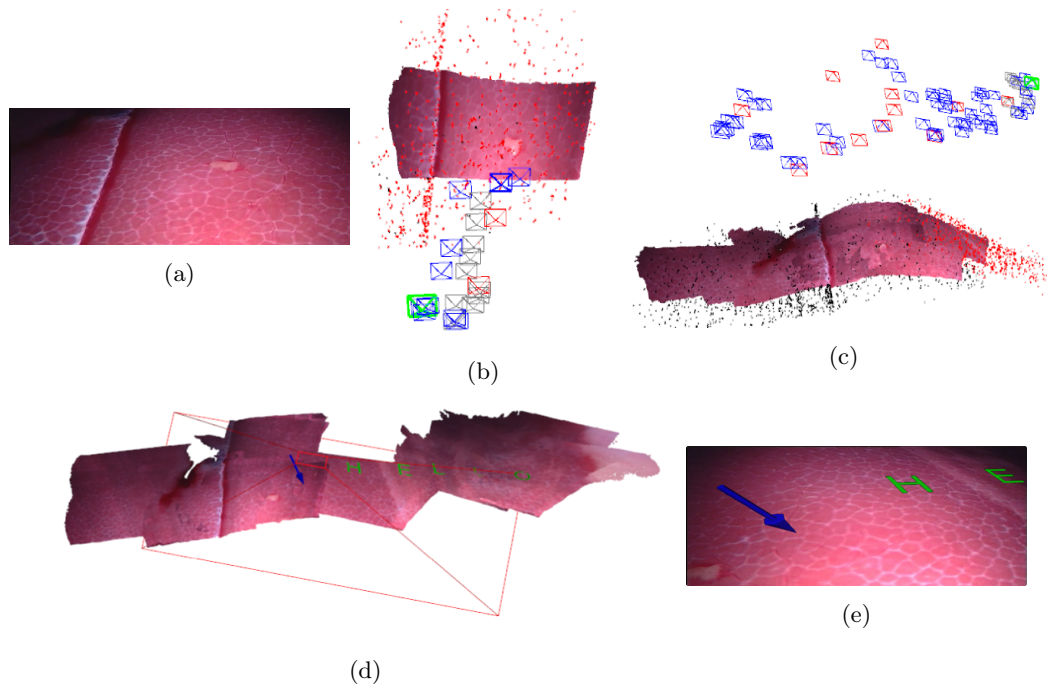


Figure 3: Visualización del endoscopio y de la reconstrucción densa incremental. (a) muestra típica imagen de una secuencia exploratoria. (b, d) Reconstrucciones densas e incrementales de la superficie del hígado (e) Visualización AR para una pose estimada de la cámara endoscópica (frustum rojo) en (d).

# Bibliography

- [1] Nader Mahmoud, Óscar G. Grasa, Stéphane A. Nicolau, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. On-patient see-through augmented reality based on visual slam. *International Journal of Computer Assisted Radiology and Surgery*, 12(1):1–11, Jan 2017.
- [2] Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In *Computer-Assisted and Robotic Endoscopy*, pages 72–83, Cham, 2017. Springer International Publishing.
- [3] Nader Mahmoud, Alexandre Hostettler, Toby Collins, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Slam based quasi dense reconstruction for minimally invasive surgery scenes. In *ICRA 2017 C4 Surgical Robots: Compliant, Continuum, Cognitive, and Collaborative*. *arXiv:1705.09107*, 2017.
- [4] Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Quasi-dense reconstruction from monocular laparoscopic video. In *Surgetica conference*, 2017.
- [5] Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Live tracking and dense reconstruction for hand-held monocular endoscopy. *IEEE Trans. on Medical Imaging (Submitted)*, 2018.

# Localisation et Cartographie Simultanées par Vision Monoculaire pour la Réalité Médicale Augmentée

Nader Mahmoud Elshahat Elsayed ALI<sup>1,3</sup>,  
**Directeurs:** Prof. José María Martínez MONTIEL<sup>2</sup>, and Prof. Christophe DOIGNON<sup>3</sup>

<sup>1</sup>*IRCAD (Institut de Recherche contre les Cancers de l'Appareil Digestif), France.*

<sup>2</sup>*Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain.*

<sup>3</sup>*ICube (UMR 7357 CNRS), Université de Strasbourg, France.*

April, 2018

## Résumé de Thèse

La chirurgie mini-invasive (CMI) a connu un gain de popularité très important au cours des deux dernières décennies. Lors de telles interventions chirurgicales une caméra endoscopique est introduite dans la cavité abdominale à travers de petites incisions effectuées sur la peau du patient, ce qui permet d'observer les structures internes de ce dernier, affichées sur un moniteur dans la salle d'opération. Les avantages de la CMI par rapport à la chirurgie traditionnelle (ouverte) sont nombreux : un séjour hospitalier plus court, de plus petites cicatrices (car seules des incisions de quelques millimètres sont nécessaires), moins de saignements, et de ce fait, un risque de traumatisme post-chirurgical bien moindre. Néanmoins, bien que de nombreux bénéfices de la CMI soient indéniables pour le patient, le geste chirurgical à exécuter est sensiblement plus difficile, et nécessite une grande expérience. Les principales difficultés rencontrées sont les suivantes : 1) la coordination œil-main dans une scène 3D observée sur un affichage 2D avec un point de vue différent des yeux du chirurgien. Il convient alors de surmonter le réflexe naturel pour diriger les yeux sur l'activité des mains; 2) la perte de la vision directe. En effet, le chirurgien ne regarde que des images 2D sur le moniteur et cela peut provoquer des perceptions erronées, en particulier la perception de la profondeur qui est cruciale pour évaluer correctement la relation spatiale entre les tissus; 3) le champ de vision (FOV) fourni par la caméra endoscopique est sensiblement plus petit que celui qu'offre la vision directe lors d'un geste par chirurgie ouverte; 4) la vision endoscopique ne fournit pas d'informations sur les structures anatomiques critiques telles que les tumeurs et les vaisseaux puisqu'ils sont localisés sous la surface des organes. Par conséquent, la CMI peut s'avérer particulièrement fastidieuse et augmente significativement la durée des opérations.

Pour surmonter les contraintes inhérentes aux interventions par la CMI, cette thèse se concentre sur l'étude et l'apport de la localisation et de la cartographie dense et simultanées par la vision monoculaire (en anglais *SLAM - Simultaneous Localisation and Mapping*). Précisément, dans le contexte d'une intervention par la CMI, les deux étapes fondamentales: 1) *la reconstruction 3D intra-opératoire dense* du champ opératoire et 2) *la localisation de l'endoscope* dans la cavité abdominale du patient, doivent être réalisées en temps réel. Le SLAM, qui est un sujet populaire en robotique et en vision par ordinateur, est une approche qui permet de construire et de mettre à jour une reconstruction 3D (appelée aussi cartographie ou *map*) d'un environnement inconnu tout en réalisant dans le même temps un suivi de la pose relative de la caméra par rapport à cette carte. Ces informations sur la localisation et la cartographie sont indispensables dans la perspective d'offrir une assistance par le guidage intra-opératoire pendant une intervention par la CMI. La représentation de la scène chirurgicale 3D dense compense le problème lié à la limitation du champ de vision et elle améliore significativement la perception de la profondeur que peut avoir le chirurgien au cours du geste. En ajoutant à cela l'estimation en temps réel de la pose relative de l'endoscope, nous disposons alors de tous les ingrédients utiles à la conception d'un système complet de guidage par la réalité augmentée (RA).

La RA permet de superposer à l'image endoscopique du patient un modèle géométrique préopératoire, véritable clone virtuel du patient. C'est du moins une des multiples manières de restituer des informations utiles pour le chirurgien. A cette fin, des clones virtuels sont créés à partir de différentes modalités d'imagerie telles que la tomodensitométrie (CT) ou la résonance magnétique (MRI). De cette manière, le chirurgien a la sensation de voir son patient en transparence. La RA peut fournir une information décisive durant la CMI comme par exemple l'identification peropératoire de structures d'intérêt non visibles (tumeurs, vaisseaux et nerfs,...). Cela permet au chirurgien de ne plus devoir adapter mentalement les informations extraites des images médicales (CT/MRI) à la scène, ou bien de le guider durant les résections en affichant les trajectoires de coupe et les marges préalablement planifiées sur les modèles préopératoires, en rendant sûr et optimal les placements des trocars, également préalablement planifiés sur un modèle préopératoire.

Le travail préliminaire que nous avons présenté dans [1], a introduit une approche fondée sur le SLAM qui permet de visualiser avec précision les modèles préopératoires intraopératoires sur la peau du patient, en utilisant seulement un Tablet-PC. Dans ce travail, une version non dense (dite clairsemée ou éparse) du SLAM visuel est utilisée pour localiser de manière robuste la position relative de la caméra du Tablet-PC par rapport au corps du patient. Cette version fonctionne sur des appareils mobiles et est destinée à la perception de scènes de petites tailles. En vue de l'utiliser dans la salle d'opération par le personnel médical, nous avons développé un ensemble de traitements, allant de l'enregistrement à la visualisation, et qui nécessite des interactions minimales de la part du personnel médical. En effet, les interactions avec le chirurgien sont réduites à l'identification de 4 à 6 références anatomiques au début de la procédure qui sont utilisées pour effectuer l'enregistrement des données préopératoires. Contrairement aux systèmes existants de visualisation par RA similaires, le système proposé ici effectue aussi le suivi de la caméra, la reconstruction peu dense de la scène observée, l'enregistrement et le rendu visuel par la RA en temps réel. De plus, il est robuste à un fort taux d'occultations de la scène et ne nécessite pas de dispositifs de suivi externes ni de repères (marqueurs) artificiels sur la peau du patient pour localiser la caméra. Ce système a été rigoureusement évalué dans une série d'expériences. Premièrement, la précision géométrique a été évaluée au moyen de plusieurs expériences avec des marqueurs ajoutés pour disposer d'une vérité terrain; les premières expériences ont été effectuées *in vivo* sur des cochons, les secondes sur un fantôme (ou mannequin). Une évaluation par le personnel médical et une estimation du temps de calcul ont été réalisées avec deux volontaires, chacun d'entre eux reposant sur la table pendant que le praticien tenait le Tablet-PC et se déplaçait tout autour. La Figure 1 montre une visualisation par la RA *in vivo* de données préopératoires sur un volontaire et sur un cochon.

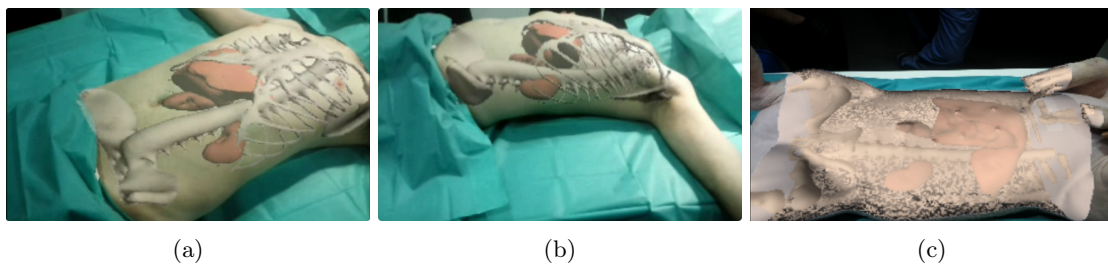


Figure 1: Visualisation par réalité augmentée (RA) de données préopératoires, proposée dans [1]. (a-b) Superposition des os, foie, reins gauche et droit sur un volontaire, pour deux points de vue. (c) Visualisation par la RA sur un cochon.

Dans un deuxième temps, nous nous sommes intéressés aux images endoscopiques. Les images endoscopiques sont très difficiles à exploiter pour calculer une représentation dense de la scène et effectuer le suivi de l'endoscope. En effet, parmi les difficultés à surmonter on peut citer le fait que de vastes régions à reconstruire sont très faiblement texturées, que l'illumination des surfaces en question varie énormément pour différentes prises de vue, avec parfois des zones spéculaires sur l'image (ceci est dû principalement à la source de lumière utilisée qui, directement fixée à l'extrémité de l'endoscope, est dirigée vers la scène). De plus, les recalages entre les vues et la scène sont fortement perturbés par les déformations des tissus et les mouvements brusques de l'endoscope dans la cavité abdominale. Nous apportons ici notre première contribution à l'exploitation du SLAM visuel dans le domaine médical. En utilisant un réglage approprié des différents paramètres

du système actuel, les performances de la partie suivi sont excellentes, comme en attestent les résultats communiqués dans [2]. Cependant, nous avons été confronté à une limitation de la qualité de la partie cartographie en terme de densité. En effet, la reconstruction de la scène chirurgicale contient très peu de points 3D en raison du faible nombre d'amers fiables dans la scène (Fig. 2(b)). La faible densité de la carte résultante empêche son utilisation pour d'autres tâches que la localisation de l'endoscope dans la cavité abdominale. Ainsi, notre deuxième contribution a consisté à proposer un algorithme de reconstruction qui améliore significativement la densité de la carte et permet une reconstruction robuste de la scène chirurgicale avec une précision de 4,9 mm [3, 4] (Fig. 2(c)). Pour cela, nous exploitons la phase d'exploration initiale de la cavité abdominale qui est typiquement effectuée par le chirurgien avant l'intervention, et acquérons un ensemble d'images clés sélectionnées. Ces images clés sont utilisées pour estimer la pose relative et construire la cartographie à l'aide d'une technique d'optimisation par ajustement de faisceaux (*Bundle Adjustment* - *BA*). Après la phase d'exploration initiale, parmi les images clés acquises, celles pour lesquelles des correspondances ont été trouvées en utilisant un algorithme de *patch-matching* sans amer, sont associées pour former des couples d'images stéréo. Grâce au suivi robuste de l'endoscope et à la représentation de la scène dense obtenue, le calcul de la visualisation (par RA intraopératoire sans marqueur) des structures internes du foie (Fig. 2 (d)) est exécuté en temps réel. La seule exigence est de transformer les modèles intra-opératoires ou préopératoires en une reconstruction dense (Fig. 2(e)). Une fois initialisée, la visualisation par RA est maintenue, pendant que les poses relatives sont estimées en continu pendant que l'endoscope explore la cavité abdominale (Fig. 2 (f)).

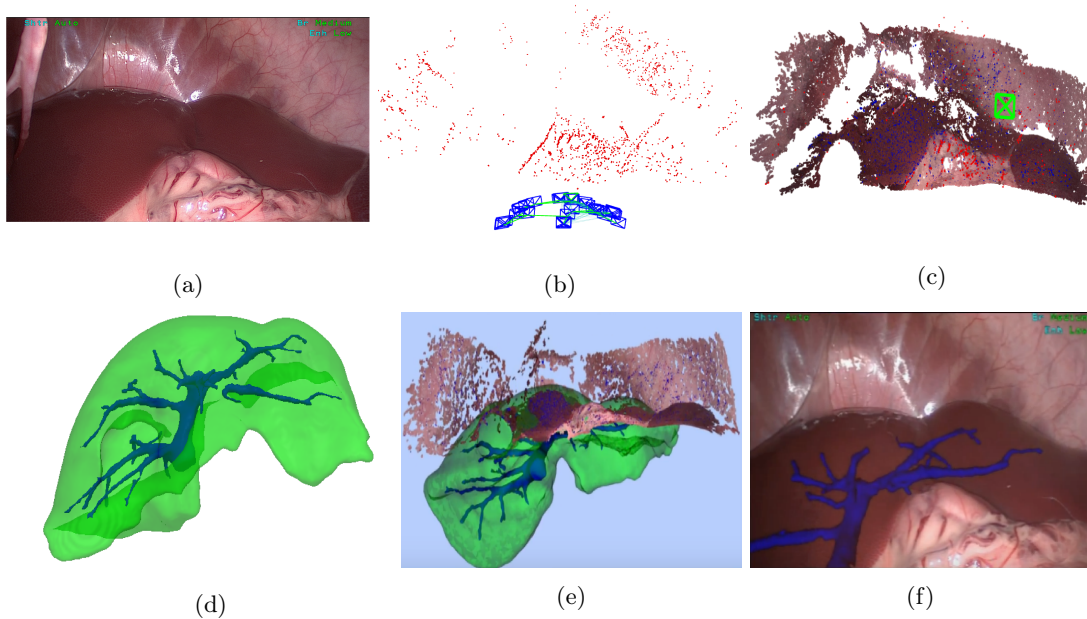


Figure 2: Reconstruction de la scène chirurgicale. (a) Une image (échantillon) de la séquence. (b) Reconstruction clairsemée par SLAM visuel. (c) Reconstruction dense par SLAM visuel. (d) Modèles de veines hépatiques segmentés semi-automatiquement à partir d'images CT. (e) Recalage du modèle intra-opératoire du foie avec une reconstruction dense. (f) Visualisation par RA sans marqueur de la veine hépatique.

Dans [5], nous avons proposé une solution toujours basée sur le SLAM visuel et la reconstruction dense en temps réel, capable de faire face aux défis de l'endoscopie et qui a été appliqué avec succès à la laparoscopie *in vivo*. Le système proposé prolonge notre travail précédent [3, 4], mais en empruntant une approche de type multi-vue, dense et stéréoscopique pour la récupération de la géométrie de la scène. Le système initial est amélioré tant du point de vue technique que méthodologique de plusieurs façons. Tout d'abord, une implémentation spécifique s'appuyant sur la programmation multitâche (*multithreading*) est mis en oeuvre : un nouveau fil d'exécution (*thread*) est ajouté au système en place et réalise la reconstruction dense de scène. Il s'exécute en direct et en parallèle des autres fils d'exécution de suivi (*tracking thread*) et de cartographie (*mapping*

*SLAM thread*). Ceci évite de devoir attendre la fin de l’exploration de la cavité abdominale pour débiter le traitement dit de *densification*, autrement dit d’affiner la reconstruction dense. Nous présentons ici une nouvelle technique qui permet de réduire significativement le temps de calcul de la tâche de densification de la reconstruction. Pour cela, nous sélectionnons parmi toutes les images de la vidéo utilisée, un sous-ensemble d’images représentatif de notre scène, appelé ensemble des images clés. Ensuite, pour chacune de ces images clés, une carte de profondeur est calculée en utilisant les images contenues dans une petite fenêtre temporelle, juste avant et juste après cette image clé (appelé *groupe de trames*). Ces cartes de profondeurs sont ensuite fusionnées pour créer une unique reconstruction. La fenêtre temporelle est automatiquement ajustée afin que la carte de profondeur en question puisse être mise à jour de manière robuste : typiquement, si la caméra bouge lentement (ce qui se traduira par une parallaxe inter-image faible), la fenêtre est agrandie afin que la carte de profondeur puisse être calculée correctement. Grâce à cela, on s’assure que le changement de point de vue de la caméra est suffisant pour ne pas générer des erreurs trop importantes dans le calcul de la profondeur. Pour calculer la carte de profondeur, nous utilisons une méthode variationnelle qui est robuste aux forts changements d’illumination. Cette approche est très différente des travaux précédents sur la densification utilisant le SLAM, qui supposent qu’il n’y a aucun changement dans la luminosité de chaque pixel au cours du temps. De plus, la méthode variationnelle que nous utilisons permet d’appréhender les variations de textures et de reconstruire des cartes de profondeurs présentant des discontinuités, ce qui se produit fréquemment dans les scènes observées par laparoscopie où les organes et les outils chirurgicaux peuvent être occultés les uns par les autres. Notre chaîne de traitements fournit une reconstruction globale et cohérente de la scène chirurgicale en fusionnant et en alignant les cartes de profondeur des images clés en ligne. Le système complet a été validé expérimentalement et évalué sur des séquences vidéo de la cavité abdominale de porcs. En outre, son exécution nécessite un temps de calcul raisonnable en utilisant les processeurs graphiques récents comme unités d’exécution de calculs. De plus, nous avons effectué une comparaison avec d’autres méthodes de SLAM visuel denses et les performances que nous obtenons sont supérieures en termes de précision, de densité et de réduction du temps de calcul.

En conclusion, grâce à la sélection efficace des trames vidéo pour générer une reconstruction dense fondée sur des critères de parallaxe, la méthode proposée peut surpasser les reconstructions purement stéréoscopiques, car le groupe de trames peut fournir une plus grande parallaxe endoscopique. Les figures 3(b-d) montrent les résultats de la reconstruction dense et incrémentale de la surface du foie, utilisée pour calculer la visualisation à l’aide de la RA (Fig. 3(d)). La pose estimée de l’endoscope (*frustum* rouge dans la Fig. 3(d)) est alors utilisée pour mettre à jour l’angle de vue du modèle virtuel pour la RA (Fig. 3(e)).

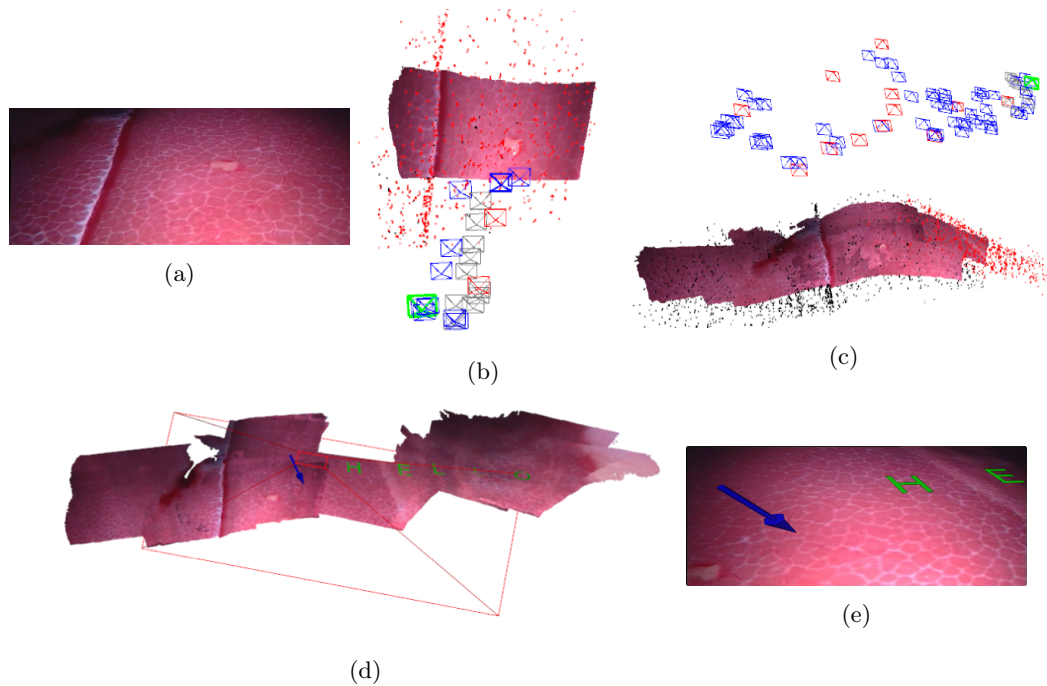


Figure 3: Suivi visuel direct de l’endoscope et reconstruction dense incrémentale. (a) Échantillon d’image d’une séquence exploratoire. (b-d) Reconstructions denses et incrémentales de la surface du foie. (e) Visualisation par RA pour une pose estimée de la caméra endoscopique (frustum rouge) en (d).

## Publications

- [1] Nader Mahmoud, Óscar G. Grasa, Stéphane A. Nicolau, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. On-patient see-through augmented reality based on visual slam. *International Journal of Computer Assisted Radiology and Surgery*, 12(1):1–11, Jan 2017.
- [2] Nader Mahmoud, Iñigo Cirauqui, Alexandre Hostettler, Christophe Doignon, Luc Soler, Jacques Marescaux, and J. M. M. Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In *Computer-Assisted and Robotic Endoscopy*, pages 72–83, Cham, 2017. Springer International Publishing.
- [3] Nader Mahmoud, Alexandre Hostettler, Toby Collins, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Slam based quasi dense reconstruction for minimally invasive surgery scenes. In *ICRA 2017 C4 Surgical Robots: Compliant, Continuum, Cognitive, and Collaborative*. *arXiv:1705.09107*, 2017.
- [4] Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Quasi-dense reconstruction from monocular laparoscopic video. In *Surgetica conference*, 2017.
- [5] Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and J. M. M. Montiel. Live tracking and dense reconstruction for hand-held monocular endoscopy. *IEEE Trans. on Medical Imaging (Submitted)*, 2018.